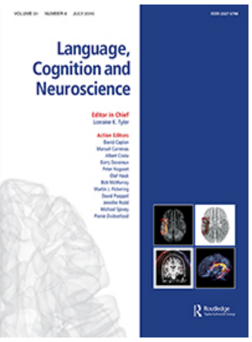

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: http://www.tandfonline.com/loi/plcp21

A sheet of coffee: an event-related brain potential study of the processing of classifier-noun sequences in English and Mandarin

Zhiying Qian & Susan M. Garnsey

To cite this article: Zhiying Qian & Susan M. Garnsey (2016) A sheet of coffee: an event-related brain potential study of the processing of classifier-noun sequences in English and Mandarin, Language, Cognition and Neuroscience, 31:6, 761-784, DOI: 10.1080/23273798.2016.1153116

To link to this article: http://dx.doi.org/10.1080/23273798.2016.1153116

Published online: 25 Apr 2016.

LANGUAGE, COGNITION AND NEUROSCIENCE, 2016
VOL. 31, NO. 6, 761–784
http://dx.doi.org/10.1080/23273798.2016.1153116

A sheet of coffee: an event-related brain potential study of the processing of classifier-noun sequences in English and Mandarin

Zhiying Qian[a,b] and Susan M. Garnsey[b,c]

[a]Department of East Asian Languages and Cultures, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [b]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [c]Departments of Psychology, Linguistics, and Neuroscience, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT
Comprehension of classifier-noun sequences was examined in separate studies in English and Mandarin by comparing event-related brain potential (ERP) responses to classifier-noun matches (*a sheet of paper*) and mismatches (*a sheet of coffee*) embedded in sentences. One goal was to determine which ERP components are sensitive to such mismatches, as a clue about the nature of the underlying combinatorial processes. Another goal was to examine effects of classifier constraint strength (*a piece of …* vs. *a sheet of …*) on anticipation of a subsequent noun. Results were similar in the two languages, which is remarkable given substantial differences between them in classifier usage. In both languages, nouns evoked larger N400s in mismatching classifier-noun sequences, suggesting that combinatorial processing was primarily semantic, and general classifiers evoked a larger sustained frontal negativity than specific classifiers starting 200 milliseconds after classifier onset, reflecting effects of constraint strength on anticipation of the upcoming noun.

ARTICLE HISTORY
Received 27 April 2015
Accepted 3 February 2016

KEYWORDS
Sentence comprehension; anticipatory processing; classifiers; N400; sustained frontal negativity

Introduction

Classifiers are words or morphemes that classify nouns by some feature of their referent, such as shape or duration. In English, classifiers are words that are used primarily to specify counting units when mass nouns are quantified. For example, the phrase *three grains of rice* includes the classifier *grain*, specifying the countable units of the mass noun *rice*. Different mass nouns require different classifiers that match some property of their meaning, such as *sheet* in *sheet of paper*, or *strand* in *strand of hair*.

Classifiers are used much more widely in some languages than they are in English. In Mandarin, for example, a classifier is required whenever any noun, including count nouns, is specified or quantified, as illustrated below in (1). As in English, Mandarin classifiers must be congruent with some aspect of the meaning of the noun they modify. For instance, *kē* is a classifier for small round solid inanimate things, and it is used with *pearl*, *candy*, *bean* and *grape*. It would be incorrect to use *kē* with *goldfish* as in (2) below, since *goldfish* instead requires *tiáo* for long soft bendable things. It is also ungrammatical if there is no classifier, as illustrated in (3) below.

(1) 鱼缸 里 有 两 条 金鱼。
Fish tank inside has two **tiao-CL** goldfish
There are two goldfish in the fish tank.
(2) *鱼缸 里 有 两 颗 金鱼。
Fish tank inside has two **ke-CL** goldfish
There are two goldfish in the fish tank.
(3) *鱼缸 里 有 两 金鱼。
Fish tank inside has two goldfish
There are two goldfish in the fish tank.

(In the English translations in (1) and (2), there is no translation of the classifier because *goldfish* is a count noun that does not take a classifier in English.)

There is ongoing debate among linguists and psycholinguists about the nature of classifiers in Mandarin and other similar languages and about how they are processed. Some linguists argue that classifiers are semantic units that set selectional restrictions for nouns they can modify (e.g. Aikhenvald, 2000; Croft, 1994; Huang & Ahrens, 2003; Wu & Bodomo, 2009). Classifiers also sometimes contribute additional meaning to a noun phrase, as in *kuài* in *yí kuài zhū ròu* (*a lump of pork*) and *piàn* in *yí piàn zhū ròu* (*a slice of pork*), where the classifiers *kuài* and *piàn* indicate the shape of the pork. In other cases

CONTACT Zhiying Qian zqian3@illinois.edu

such as *yí lì mǐfàn* (*a grain of rice*), the classifier *lì* may add to the meaning by focusing on the individual elements of the mass noun.

In contrast, other linguists have argued that classifiers in Mandarin and other similar languages are best characterised as syntactic entities, since they are syntactically obligatory for all quantified and specified nouns (Li & Thompson, 1981). It has been suggested that one function of classifiers in Mandarin is to partially compensate for the absence of number-marking morphology (e.g. Gebhardt, 2011; Greenberg, 1972; Klein, Carlson, Li, Jaeger, & Tanenhaus, 2012; Krifka, 1995; Lehman, 1979; Ritchie, 1971). When a noun is quantified, the quantifier itself provides information about number, but for nouns that are specified but not quantified (e.g. *this book*), it is only the classifier that provides information about number. For example, the classifier *běn* (*book-like*) in the phrase *zhè běn shū* (*this book-like book*) provides the information that the noun is singular, compared to a bare *shū*, which could be either singular or plural. This example also illustrates another argument for considering classifiers to be primarily syntactic entities in Mandarin, since the meaning of the classifier *běn* is completely redundant with the meaning of the noun *shū*. Another proposal is that the count/mass distinction is syntactically realised in Mandarin by using count classifiers for count nouns and mass classifiers (i.e. measuring units) for mass nouns (Cheng & Sybesma, 1999, 2005; Chierchia, 1998; Gebhardt, 2011). On this view, Mandarin classifiers sometimes function similarly to the way definite determiners and number morphology do in English.

One factor that could be important in how classifiers are processed is their lexical status in different languages. English classifiers are generally nouns that can stand alone and be used in other ways (e.g. *There was enough grain stored to last the winter*), and there is not the same debate among linguists about whether English classifier-noun integration is better characterised as semantic or syntactic. In contrast, Mandarin classifiers are typically bound morphemes that can only combine with numbers or demonstrative determiners and can rarely be used alone as nouns. Thus, it might be better to think of the requirement that classifiers and nouns match as a kind of agreement in Mandarin, like gender agreement in languages that mark gender, rather than as a more general requirement for semantic congruity like using an adjective that is appropriate for the noun it modifies. If it is indeed more appropriate to think of classifier-noun match in Mandarin as a kind of agreement, it would be the only such instance in Mandarin. Because words are not marked morphologically in Mandarin for number, person, gender, or argument roles, there can be no agreement between words on such features. Thus, agreement between words in sentences is a much rarer phenomenon in Mandarin than in morphologically rich languages.

In languages that have morphosyntactic agreement, it has been found to be processed differently from the integration of the meanings of content words into a sentence-level interpretation. An important tool demonstrating this has been event-related brain potentials (ERPs), which have proved to be informative because different ERP components are sensitive to different aspects of language processing. The amplitude of the N400 component, a negative-going deflection peaking approximately 400 milliseconds (ms) after the onset of a word, is sensitive to how easy it is to process the word's meaning in context (Kutas & Hillyard, 1980). N400 is part of the response to every word and its amplitude varies depending on how much information is retrieved and how easy it is to integrate it with the context (for a review, see Kutas & Federmeier, 2011). In contrast, when words in a sentence that are required to agree fail to do so, the result is instead typically an increase in the amplitude of P600, a positive-going component peaking somewhere around 600 ms after the onset of the word signalling a problem. P600 is sensitive to a variety of aspects of form and structure processing in sentences besides agreement (Friederici, Pfeifer, & Hahne, 1993; Gouvea, Phillips, Kazanina, & Poeppel, 2010; Hagoort, Brown, & Groothusen, 1993; Kaan, Harris, Gibson, & Holcomb, 2000; Kaan & Swaab, 2003; Neville, Nicol, Barss, Forster, & Garrett, 1991; Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994). Recent findings of "semantic P600" effects have complicated the picture somewhat (Bornkessel-Schlesewsky & Schlesewsky, 2008; Chow & Phillips, 2013; Hoecks, Stowe, & Doedens, 2004; Kim & Osterhout, 2005; Kolk, Chwilla, Van Herten, & Oor, 2003; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Van Herten, Kolk, & Chwilla, 2005), but it is still appropriate to describe the P600 as sensitive to structure processing and its consequences, which can include situations where there is conflict between structure-based and meaning-based interpretations (Kuperberg, 2007).

A third family of ERP components sensitive to aspects of language processing are often called Anterior Negativities because they are negative-going deflections that are maximal at anterior scalp sites. They sometimes precede P600 effects but other times occur on their own. Sometimes the negativity is larger over the left hemisphere and called a left anterior negativity (LAN) (Friederici et al., 1993; Kluender & Kutas, 1993). One variety of anterior negativity is similar in onset and duration to the N400 but with a more frontal scalp distribution, while other varieties begin earlier or later and/or

Language, Cognition and Neuroscience 2016.31:761-784.

persist longer. It is not yet clear whether and how all of the various anterior negativities are related to one another, since a wide range of phenomena have been found to elicit them, including morphosyntactic agreement violations (Coulson, King, & Kutas, 1998; Gunter, Friederici, & Schriefers, 2000; Osterhout & Mobley, 1995), word concreteness and imageability (Gullick, Mitra, & Coch, 2013; Holcomb, Kounios, Anderson, & West, 1999; Kounios & Holcomb, 1994; Lee & Federmeier, 2008; Zhang, Guo, Ding, & Wang, 2006), working memory load related to complex structure (King & Kutas, 1995; Kluender & Kutas, 1993; Weckerly & Kutas, 1999), "frame-shifting" in processing non-literal language (Coulson & Kutas, 2001), and lexical (Lee & Federmeier, 2009; Wlotko & Federmeier, 2011, 2012) or referential (Nieuwland, Otten, & Van Berkum, 2007) ambiguity. Other sustained anterior negativities have been observed in domains other than language and have been found to be related to anticipation of an upcoming event with known timing, including the stimulus preceding negativity (SPN; see Brunia, van Boxtel, & Boecker, 2012 for a review) and contingent negative variation (see Tecce & Cattanach, 1993 for a review). It is not yet clear whether and how these components are related to the anterior negativities that have been found in language studies. They are included here because they might be relevant for our studies, which examine the degree to which particular nouns can be anticipated following classifiers in sentences with words presented at a fixed rate.

The interpretation of the various anterior negativities has become more complicated recently because it appears that an anterior negativity can sometimes result from superposition of partially overlapping N400 and P600 effects that cancel each other out to varying degrees at different scalp sites, depending on the amplitude of each component (Tanner, 2015). Yet another complication is that there are individual differences in ERP responses to the same stimuli among both native and non-native speakers, with some people producing P600-dominant responses and others N400-dominant responses (Osterhout, 1997; Tanner, Inoue, & Osterhout, 2014; Tanner & Van Hell, 2014). It has also been shown that the same syntactic agreement violations yield different responses depending on whether they are in word pairs or sentences (Barber & Carreiras, 2005; Münte, Heinze, & Mangun, 1993). Finally, it has been demonstrated that task can affect which component is dominant in the ERP responses (Hahne & Friederici, 2002; Oines & Kim, 2014). However, while it is clear that there is not a simple correspondence between N400 and meaning processing on the one hand and P600 and/or anterior negativities and structure processing on the other, determining which ERP components are sensitive to the processing of particular kinds of words in particular kinds of contexts still provides useful evidence about language comprehension, although care must be taken when drawing conclusions about the types of processes underlying particular ERP components.

In English, classifiers are nouns that can stand alone and be used in other ways than specifying counting units for mass nouns. Given what is known so far about language-sensitive ERP components, it seems highly likely that a noun that does not match a classifier in English would evoke an increase in N400 amplitude, just like a noun that does not fit with an adjective preceding it. However, it is less clear what to expect in languages like Mandarin. In languages with morphosyntactic agreement, it is usually the P600 that has been found to be sensitive to violations, even when substantial semantic/pragmatic processing is required, such as gender agreement between a pronoun and an antecedent naming an occupation that is stereotypically associated with one gender (e.g. *nurse* and *she*; Osterhout, Bersick, & McLaughlin, 1997; Osterhout & Mobley, 1995; see however Severens, Jansma, & Hartsuiker, 2008 and Nieuwland, Martin, & Carreiras, 2013 for N400 effects in response to some kinds of agreement violation). If Mandarin classifier-noun sequences are processed similarly to adjective-noun sequences, it is likely that N400 would be larger for violations, similar to the expectation for English. If, however, Mandarin classifier-noun sequences are processed more like morphosyntactic agreement in other languages, P600 might be the component most likely to respond to violations.

A potentially important property of classifiers is that they vary in how strongly they constrain what nouns can follow them. Some classifiers can be used with only a very few nouns while others can take many types of nouns. For instance, the Mandarin classifier *zhăn* is used almost exclusively with lamps whereas *kē* can be used with anything that is small, solid, and round. Highly specific classifiers like *zhăn* (lamp-like) make the subsequent noun much more predictable than general ones like *kē* (small, solid, and round).

Debate about whether the language comprehension system actively makes predictions about likely upcoming words and/or their features is longstanding. Evidence has been accumulating in support of prediction of some kinds of information under some circumstances (e.g. Altmann & Kamide, 1999; Arai & Keller, 2013; Chow, Smith, Lau, & Phillips, 2015; DeLong, Urbach, & Kutas, 2005; Dikker & Pylkkänen, 2013; Dikker, Rabagliati, & Pylkkänen, 2009; Federmeier, 2007; Federmeier & Kutas, 1999; Federmeier, Kutas, & Schul, 2010; Fruchter, Linzen, Westerlund, & Marantz, 2015; Kaiser & Trueswell, 2004; Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003; Kim & Gilley, 2013; Kim & Lai, 2012; Kim, Oines, & Sikos, 2015; Kuperberg & Jaeger, 2016; Lewis & Bastiaansen, 2015; Lewis, Wang, & Bastiaansen, 2015; Szewczyk & Schriefers, 2015; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004; Wlotko & Federmeier, 2015). For instance, DeLong et al. (2005) asked participants to read sentences like *The day was breezy so the boy went out to fly …* , for which there was a highly predictable completion (*a kite),* as well as plausible but less predictable completions (*an airplane*). They found a larger N400 in response to the determiner *an* compared to *a*, suggesting that the highly predictable noun *kite* was pre-activated prior to its appearance. Since the effects were seen on the determiner preceding the highly predictable noun, it is clear that properties of the noun (i.e. whether it began with a vowel) were predicted before it appeared. Van Berkum et al. (2005) and Wicha et al. (2004) have also found evidence for anticipation of specific words using grammatical gender in Dutch and Spanish, respectively, though the ERP components involved have differed across studies. Fruchter et al. (2015) recently reported that the frequency of occurrence of a highly predictable noun (made predictable by the adjective preceding it) affected MEG responses before the noun appeared, and Szewczyk and Schriefers (2015) found ERP effects in Polish when discourse context made the animacy of a noun highly predictable but an adjective appeared with a mismatching animacy marker. In all of these studies, properties of a predictable word affected responses before that word appeared, showing that those properties were preactivated before the predicted word appeared.

Another ERP component that is sensitive to the predictability of visual stimuli is P2, a frontally distributed positivity peaking 200–300 ms after stimulus onset (Hillyard & Münte, 1984; Luck & Hillyard, 1994). Some recent studies have found P2 amplitude to be affected by the predictability of words in phrasal or sentential contexts (Dambacher, Rolfs, Goellner, Kliegl, & Jacobs, 2009; Federmeier & Kutas, 2002; Federmeier, Mai, & Kutas, 2005; Huang, Lee, & Federmeier, 2010; Lee, Liu, & Tsai, 2012). (See Kaan and Carlisle (2014) for similar effects for letters in alphabetic vs. random sequences.) For final words in strongly vs. weakly constraining sentence contexts (e.g. *She was suddenly called back to New York and had to take a cab to the airport*. vs. *She was glad she had bought a book since there was nothing to read at the airport*), Federmeier et al. (2005) found larger P2s in strongly constraining contexts.[1] The effect of sentential constraint on P2 amplitude was initially explained as facilitation of the processing of predictable visual features of predicted words. However, the results of subsequent studies called that interpretation into question. What seems to matter is how constraining the context preceding a word is, not whether the final word is the most predicted one (Wlotko & Federmeier, 2007, 2011). Thus, the amplitude of the P2 response to a word seems to reflect something about the state engendered by strongly constraining linguistic context preceding it, rather than whether the word matches the predictions afforded by the context.

In most of the studies finding effects of contextual constraint on P2 amplitude, strong contexts have been built up across multiple words and it is sometimes not clear exactly when during the context the target word becomes highly predictable. Classifiers provide a way to manipulate the strength of contextual constraint via a single word. (See also Fruchter et al., 2015; and Huang et al., 2010, who manipulated constraint strength using just adjectives.) There seem to have been no previous studies of classifier-noun integration in English, so the first step here is to conduct such a study. (There have been previous studies of Mandarin and Japanese classifiers, but review of those will be postponed until after the English study is presented here.) The most likely result in English is that N400 amplitude will increase when a noun mismatches the classifier preceding it. It is also possible that the strength of the constraint provided by classifiers will affect the amplitude of the N400 and/or P2 responses to the nouns. Finally, it is possible that some aspect of the response to the classifiers themselves may reflect the degree to which they constrain what can follow them.

Experiment 1

Method

Participants

Thirty native-English-speaking undergraduate students at the University of Illinois at Urbana-Champaign (17 females; mean age 19.2; range 18–22) participated for course credit. All had normal or corrected-to-normal vision, were strongly right-handed as assessed by the Edinburgh inventory (Oldfield, 1971), and had no neurological or psychiatric disorders according to self-report. Thirteen participants had left-handed relatives. All gave written informed consent. Four additional participants (three females) were run but excluded from data analysis due to excessive blinking or motion artefacts.

Materials and design

Classifier specificity was determined with a norming study, in which a separate group of 19 participants provided as many nouns as they could that fit with each of 87 classifiers. Twenty general (mean 6.4 nouns, range 1.3–14.2) and 20 specific (mean 2.2 nouns, range 0.5–5.9; $F(1, 38) > 300$, $p < .01$) classifiers were chosen based on the norming study. Mean length did not reliably differ between general (4.5) and specific classifiers (4.6; $F < 1$), nor did their frequency of occurrence (general: 45.5/million words, specific: 37.6/million words, $F < 1$; all values per million words measured from SUBTLEX$_{US}$ Corpus, Brysbaert & New, 2009). The non-reliable numerical difference in frequency between the two types of classifiers was larger than would be ideal. It was simply not possible to match the frequencies of the two types of classifiers closely because frequency is necessarily negatively correlated with specificity. General classifiers are used more often than specific ones, largely because they are used with more different nouns. In fact, closely matching the mean frequencies of the two types of classifiers would result in odd subsets of each type of classifier. It will be important to keep the frequency difference in mind when interpreting the ERP responses to the classifiers, since word frequency has been found to affect both P2 and N400 amplitude, with both components having smaller amplitude for higher frequency words (Hauk, Davis, Ford, Pulvermuller, & Marslen-Wilson, 2006; Hauk & Pulvermuller, 2004; King & Kutas, 1998; Van Petten & Kutas, 1990; Young & Rugg, 1992).

Each of the 40 classifiers was used to create three different sentence pairs, in order to have a sufficient number of experimental items (120 pairs). Within each pair, one sentence had a matching classifier and noun and the other had mismatching ones. Each of the three sentence pairs using the same classifier had different critical nouns, which were the three nouns produced most often by participants in the norming study for the classifier. Each participant saw each classifier three times, either twice in the match condition (in two different items) and once in the mismatch condition (in a different item), or vice versa. Critical nouns were matched for mean length (general: 5.3, specific: 5.5, $F < 2$, $p > .1$) and frequency of occurrence (general: 98.1, specific: 97.8, $F < 1$).

Four conditions were created by crossing two levels of classifier-noun match with two levels of classifier specificity. Each sentence pair included match and mismatch versions that differed only in the classifier (e.g. *There are several [grains/heads] of rice that did not get cooked*). The sentences in the mismatch condition were created by re-pairing nouns and classifiers, so the same classifiers and nouns were used in the match and mismatch conditions. All experimental sentences had the sentence frame: *There* + *be* + quantifier + classifier + *of* + noun + relative clause. If the classifier in the match version in a pair was general, the classifier in the mismatch version was also general, and the same was true for items with specific classifiers. Sentences were distributed over two lists such that participants saw only one member of each item pair and equal numbers of items in each condition (30). Lists were divided into five blocks of 40 trials each and classifiers did not appear more than once within a block.

It was necessary to repeat some of the critical nouns across items in order to use the nouns that fit best with the classifiers in the matching conditions. While it would have been ideal to avoid repeating any of the critical nouns, since repetition is known to reduce N400 amplitude (Rugg, 1990; Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991), it was deemed more important here to use the nouns that fit best with each classifier in the matching condition, and doing so required repeating some nouns. In all, 20 critical nouns appeared in more than one item, approximately equally distributed across conditions. With two exceptions, at least 10 trials intervened between two instances of the same noun. For the exceptions (one in the general condition and one in the specific condition), one noun was repeated after nine intervening trials and another after five intervening trials.

Eighty distractors were added for a total of 200 trials/list. Four types of distractors were included: (1) sentences starting with *there* + *be* + quantifier that did not contain classifiers (20; e.g. *There are many people signed up for the psychology class*), (2) sentences with classifiers and matching nouns used in various sentence positions (20; e.g. *The woman is wearing a nice pair of earrings*), (3) sentences using potential classifiers as nouns instead (20; e.g. *This team won the game*), and (4) sentences with mismatching classifiers and nouns where it was the noun, rather than the classifier, that did not fit well with the rest of the sentence (20). Half of these had the same structure as the experimental items (e.g. *There was a flock of bread that rested on the shore*), and half placed the classified noun in other sentence positions (e.g. *The baker made four batches of fresh bikes*). The distractors were intended to prevent participants from expecting that what would follow *There* + *be* + quantifier would always be a classifier, what positions classifiers might appear in, and when there was mismatch at the noun following the classifier whether the rest of the sentence would be consistent with the noun or the classifier. Each list contained 120 fully acceptable sentences and 80 sentences with classifier-noun mismatches. All sentences were followed by comprehension questions that never specifically probed the comprehension of the classifier, half with *yes* and half with *no* responses.

Procedure

Participants were seated comfortably in a dimly lit sound-attenuating booth in front of a 23-inch LCD monitor. Each trial began with a fixation point in the centre of the screen for 500 ms. Because eye movements cause artefacts that contaminate the EEG signal, sentences were presented word-by-word at the centre of the screen in 24-point white Arial font on a black background, at a rate of 500 ms per word (300 ms text, 200 ms blank screen). After each sentence, participants responded to a comprehension question by pressing one of two buttons on a Cedrus RB-830 response box and received immediate accuracy feedback (e.g. *Was all of the rice cooked?* after *There were several grains of rice that did not get cooked*). Stimulus presentation was controlled by the Presentation® software package (www.neurobs.com). Each list was divided into five blocks, each beginning with four distractor items. Participants were given a short break after each block and instructed to minimise blinking and body movement during the sentences but to blink and move between trials when necessary. A practice block of nine trials was given at the beginning. It took about 45 min to complete 200 trials and the entire session lasted about 2 h.

EEG recording and data analysis

Continuous EEG was recorded from 27 Ag/AgCl sintered electrodes placed in an elastic cap (EasyCap, 10–10 system; Chatrian, 1985), referenced online to the left mastoid and re-referenced offline to the average of left and right mastoids: midline: Fz, Cz, Pz; lateral: AF3/4, F3/4, F7/8, FT7/8, FC3/4, C3/4, T3/4, CP3/4, T5/T6, P3/4, P5/6, PO7/8. Eye blinks and eye movements were monitored with electrodes above and below the right eye and at the outer canthi of both eyes. EEG and EOG recordings were amplified by a Grass Model 12 amplifier and sampled at a frequency of 200 Hz. A 0.01–30 Hz analogue bandpass filter was applied during online recording and a 0.1 Hz high-pass digital filter was applied offline. Impedances were maintained below 5kΩ.

Epochs were extracted from the continuous waveforms from 100 ms before the onset of the classifier through 2100 ms later, capturing the responses to the classifier, *of*, and the critical noun. Trials contaminated with artefacts during this epoch were rejected using the ERPLAB toolbox (Lopez-Calderon & Luck, 2014). Blinks and eye movements were detected using a moving window peak-to-peak function with a threshold of 50 µV on the EOG channels, and non-ocular artefacts were identified using the moving window peak-to-peak function applied to the EEG channels, with individualised thresholds determined by visual inspection of each participant's data. Epochs contaminated with artefacts were discarded, leading to an average loss of 13% of the data, which did not differ across conditions.

ERP waveforms were analysed using two different baselines: (1) 100 ms before classifier onset and (2) 100 ms before critical noun onset. The first baseline allowed examination of differences between conditions triggered by the classifier, and whether such differences persisted across the noun, while the second baseline allowed examination of differences triggered by the noun itself. Mean amplitudes were calculated for each channel in each condition for each participant for three time windows intended to capture the P2, N400, and P600 ERP components. Consistent with previous studies (e.g. Huang et al., 2010; Wlotko & Federmeier, 2007), a 50 ms time interval surrounding the P2 peak in the grand mean waveforms was chosen for the P2 measurement window. The classifier's P2 peaked at about 255 ms, so the time window used to capture P2 effects was 230–280 ms. Conventional time windows were used for the N400 (350–550 ms) and P600 (600–900 ms) components. Window mean amplitudes were submitted to repeated-measures analyses of variance (ANOVAs). One set of analyses included all electrodes and another included just midline electrodes. The ANOVA including all electrodes had four within-participant factors: two levels of classifier-noun match (match, mismatch), two levels of classifier specificity (general, specific), three levels of electrode site anteriority (frontal, central, posterior), and three levels of electrode site laterality (left, midline, right). The ANOVA including just midline electrodes (Fz, Cz, Pz) consisted of the same within-participant factors except that there was no laterality factor. When interactions with electrode site in the omnibus ANOVAs motivated further analysis, analyses were conducted on six regions of interest (ROIs), each comprising four electrodes: left anterior (AF3, F3, F7, FT7), right anterior (AF4, F4, F8, FT8), left central (FC3, C3, CP3, T3), right central (FC4, C4, CP4, T4), left posterior (P3, T5, P5, PO7) and right posterior (P4, T6, P6, PO8). Analyses within ROIs included two within-participant factors: two levels of classifier-noun match (match, mismatch) and two levels of classifier specificity (general, specific). Follow-up tests are reported when ANOVAs revealed significant interactions. The Greenhouse-Geisser correction was applied wherever necessary to correct for violations of sphericity (Greenhouse & Geisser, 1959). Corrected *p*-values and original degrees of freedom are reported. Grand average ERPs were digitally low-pass filtered at 10 Hz to smooth the waveforms for display, but analyses were performed before such filtering was applied.

Results

Behavioural results

Comprehension accuracy for experimental sentences was above 89% for all participants (mean 95%), and did not differ depending on whether the classifier was specific (95%) or general (95%; $F < 1$), or whether the classifier matched the noun (96%) or not (95%; $F < 1$), nor was there any interaction between the two factors ($F < 1$).

ERP results

There were no effects of electrode site laterality in any of the analyses, so all results will be described collapsing over that factor.

Classifier

Each 2100 ms epoch included the responses to the classifier, the word *of*, and the critical noun (e.g. *grain of rice*). Figure 1 shows the grand average ERPs at all channels starting 100 ms before the onset of the classifier and continuing throughout the responses to the word *of* and the critical noun in the general match, general mismatch, specific match, and specific mismatch conditions, aligned on a 100 ms baseline prior to the onset of the classifier. Visual inspection revealed more sustained frontal negativity elicited by general classifiers than by specific classifiers starting at the classifier's P2 time window and continuing throughout the epoch. This observation was confirmed by statistical analyses.

General classifiers elicited more negativity than specific classifiers at frontal sites starting at the classifier's P2 time window, leading to a significant specificity × anteriority interaction in both the overall ($F(2, 58) = 5.5$, $p < .05$) and midline analyses ($F(2, 58) = 7.0$, $p = .01$), as well as a marginally significant main effect of specificity ($F(1, 29) = 4.2$, $p = .051$) in the midline analysis. The interaction arose because responses to general classifiers were reliably more negative than to specific classifiers in the P2 time window only for the anterior region (anterior: $F(1, 29) = 7.2$, $p = .01$; central: $F(1, 29) = 2.9$, $p = .1$; posterior: $F < 1$).

The greater negativity at anterior sites for general classifiers persisted across the classifier's subsequent N400 time window (specificity × anteriority overall: $F(2, 58) = 5.0$, $p < .05$; midline: $F(2, 58) = 6.5$, $p = .01$). The effect was somewhat weaker than in the earlier P2 time window, however, since it was reduced to marginality for the anterior region in the regional analysis (anterior: $F(1, 29) = 2.9$, $p = .1$, central and posterior: $Fs < 1$). As expected, there were no effects involving classifier-noun match in any of the classifier's time windows (all $Fs < 1.2$), since the noun had not yet appeared.

Critical noun

Pre-classifier baseline. When the response to the critical noun was baselined on the 100 ms preceding the classifier, as shown for all sites in Figure 1 and again for just the midline sites in Figure 2(a), the negativity triggered by general classifiers at anterior sites persisted throughout the response to the noun, although it weakened over time, with the specificity × anteriority interaction that was reliable in the overall analysis at the classifier becoming only marginal at the noun in all three measurement windows (P2: $F(2, 58) = 3.4$, $p < .1$; N400: $F(2, 58) = 3.2$, $p < .1$; P600: $F(2, 58) = 3.1$, $p < .1$). These marginal interactions were due to specificity effects being limited primarily to the anterior region (anterior P2: ($F(1, 29) = 7.4$, $p = .01$; anterior N400: $F(1, 29) = 6.1$, $p < .05$; anterior P600: $F(1, 29) = 3.1$, $p = < .1$; central P2: ($F(1, 29) = 4.5$, $p < .05$); central N400: $F < 1$; central P600: $F(1, 29) = 1.9$, $p > .1$; posterior, all time windows: $Fs < 1$). To summarise, general classifiers triggered greater frontal negativity than specific classifiers and that difference persisted throughout the noun's time windows. There were no reliable effects of classifier-noun match in any of the noun's time windows when the pre-classifier baseline was used, although the specificity × match interaction was marginal in the N400 window in the overall analysis ($F(2, 58) = 3.3$, $p < .1$).

Pre-noun baseline. To separate the effects triggered by the critical noun from continuing effects that began at the classifier, the waveforms were re-baselined on the 100 ms before noun onset, after which clear classifier-noun match effects emerged. The waveforms at midline sites are shown in Figure 2(b) and the ANOVA results in the rightmost columns in Table 1.

Visual inspection showed that the waveforms were more negative for the mismatch than the match conditions beginning in the noun's P2 time window and persisting throughout the rest of the epoch, with the difference maximal during the N400 time window. Both the overall and midline ANOVAs revealed a reliable main effect of classifier-noun match in the N400 window (overall: $F(1, 29) = 16.6$, $p < .01$; midline: $F(1, 29) = 20.6$, $p < .01$) and reliable or marginal effects in both the P2 (overall: $F(1, 29) = 3.5$, $p < .1$; midline: $F(1, 29) = 4.1$, $p < .1$) and P600 (overall: $F(1, 29) = 3.7$, $p < .1$; midline: $F(1, 29)$ 6.8, $p < .05$) windows. There were no longer any main effects of classifier specificity in any time window when the waveforms were baselined before the noun. However, a match × specificity interaction emerged in the noun's N400 time window in

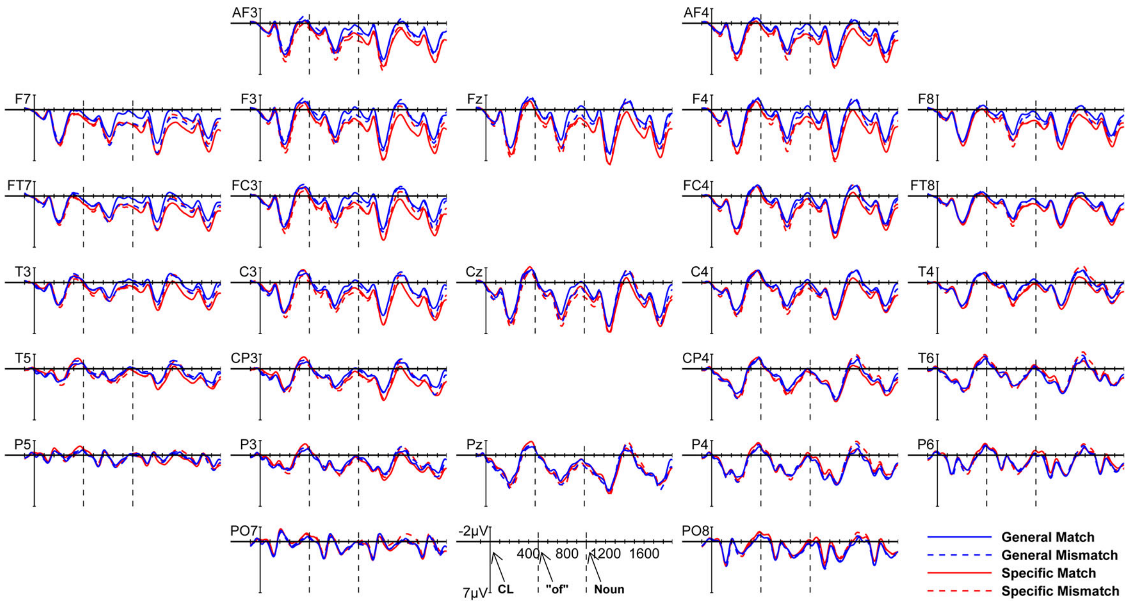

Figure 1. Grand average ERPs at all electrode sites in Experiment 1 baselined on 100 ms before classifier. Note: *Y*-axis position indicates onset of classifier.

the overall analysis (overall: *F*(1, 29) = 4.4, *p* < .05), which arose because the effect of match was larger for specific than for general classifiers during the noun's N400 window, but reliably so only in the central region (*F*(1, 29) = 6.0, *p* < .05) in the regional analysis.

Discussion

The present study investigated the processing of classifier-noun sequences in English sentences. As expected for English, the response to a noun that did not match the classifier preceding it was a larger N400. In addition, the noun's N400 was sensitive to the specificity of the classifier preceding it, showing a reliably larger mismatch effect when the classifier was specific, consistent with other studies showing that words that do not fit their sentential contexts elicit larger N400 effects when the context is more constraining (Federmeier et al., 2005; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Wlotko & Federmeier, 2012). The match effect persisted

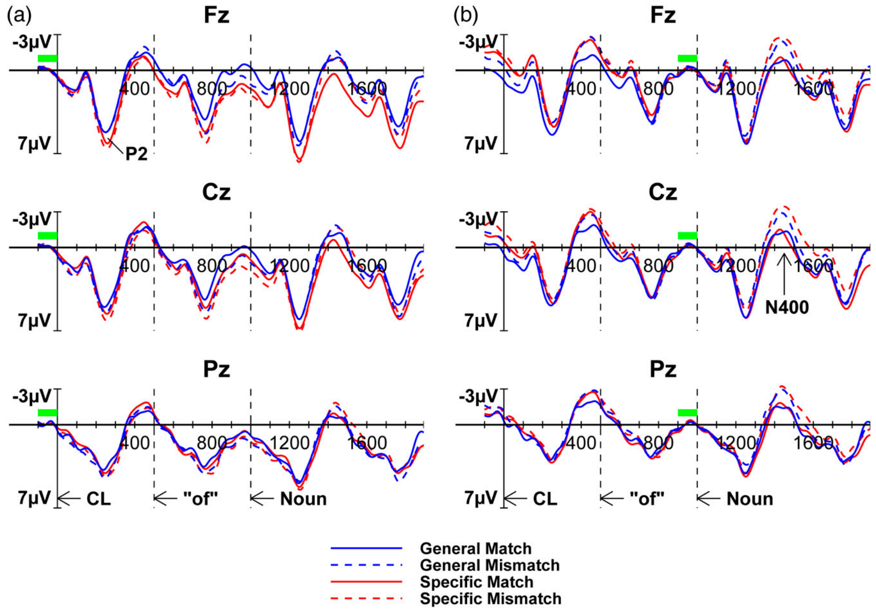

Figure 2. Grand average ERPs at midline sites Fz, Cz, and Pz in Experiment 1 baselined on (a) 100 ms before classifier and (b) 100 ms before noun. Note: Baseline interval is indicated by green bars.

Table 1. ANOVA *F*-values for Experiment 1.

	Classifier (pre-CL baseline)		Noun (pre-CL baseline)			Noun (pre-noun baseline)		
	P2	N400	P2	N400	P600	P2	N400	P600
Overall analysis								
Match (1, 29)	–	–	–	–	–	3.5!	16.6**	3.7!
Spec (1, 29)	−2.8!	–	4.1!	–	–	–	–	–
Match × Spec (1, 29)	–	–	–	3.3!	–	–	4.4*	–
Spec × Ant (2, 58)	5.5*	5.0*	3.4!	3.2!	3.1!	–	–	–
Midline analysis								
Match (1, 29)	–	–	–	–	–	4.1!	20.6**	6.8*
Spec (1, 29)	4.2!	–	4.2*	–	–	–	–	–
Match×Spec (1, 29)	–	–	–	–	–	–	–	–
Spec × Ant (2, 58)	7.0**	6.5**	–	3.6*	4.6*	–	–	–
Anterior region								
Match (1, 29)	–	–	–		–	–	7.8**	–
Spec (1, 29)	7.2*	2.9!	7.4*	6.1*	3.1!	–	–	–
Match × Spec (1, 29)	–	–	–	3.5!	–	–	–	–
Central region								
Match (1, 29)	–	–	–	–	–	3.9!	16.8**	3.4!
Spec (1, 29)	2.9!	–	4.5*	–	–	–	–	–
Match × Spec (1, 29)	–	–	–	3.2!	–	–	6.0*	2.9!
Posterior region								
Match (1, 29)	–	–	–	3.5!	–	–	15.4**	–
Spec (1, 29)	–	–	–	–	–	–	–	–
Match × Spec (1, 29)	–	–	–	–	–	–	–	–

Notes: Regional analyses were conducted on lateral electrodes, following up the overall analysis. Spec = specificity; Ant = anteriority; Lat = laterality; P2 = 230–280 ms; N400 = 350–550 ms; P600 = 600–900 ms.
**$p < .01$.
*$p < .05$.
!$.05 < p < .1$.

into the P600 window with the mismatch conditions continuing to be more negative. Note that that is in the opposite direction from what would be expected if mismatching nouns elicited a P600 effect. Thus, it appears that classifier-noun integration in English sentences is no different from general combinatory semantic processing of words in sentences.

It was also predicted that the critical noun's P2 amplitude might be affected by the specificity of the classifier preceding it, based on earlier studies showing effects of the strength of contextual constraint preceding a word on that word's P2 amplitude (Federmeier et al., 2005; Wlotko & Federmeier, 2007, 2011). What was observed instead was a sustained frontal negativity evoked by general classifiers, starting during their P2 time window and persisting throughout the rest of the measurement epoch. There are reasons to think that this result is not an instance of the enhancement of P2 amplitude by predictive contexts that has been found in some previous studies. First, it was triggered by the classifier itself, rather than by the noun following it. In contrast, P2 enhancement in most previous studies has appeared on a word that is made predictable by a highly constraining context, rather than on the constraint-providing word itself (Wlotko & Federmeier, 2007, 2011).[2] However, two recent studies have reported sustained effects of highly constraining context that began during the P2 window in the response to the constraining stimulus itself, much like ours. Chou, Huang, Lee, and Lee (2014) examined the combined effects of classifier constraint strength and classifier-noun match in Mandarin using an explicit congruity judgment task on sequentially presented word pairs, and found greater negativity beginning during the P2 time window in response to the low-constraint classifiers. They described their results as an instance of P2 amplitude enhanced by the predictive context supplied by high-constraint classifiers, followed by a slower and longer-lasting frontal negativity in the less predictive context provided by low-constraint classifiers. However, it seems more parsimonious to characterise the effect as a sustained frontal negativity in response to low-constraint classifiers that began during their P2 time window, especially given that the scalp distribution of the effect did not change over time. Sustained frontal negativities like this have been found previously in conditions where there is greater uncertainty or ambiguity, such as in non-literal language processing (Coulson & Kutas, 2001) and some kinds of lexical (Federmeier et al., 2007; Lee & Federmeier, 2009; Wlotko & Federmeier, 2012) or referential ambiguity (Nieuwland et al., 2007; Van Berkum et al., 2005). In both Chou et al.'s results in Mandarin and our results in English, a sustained frontal negativity was observed in response to less-constraining general classifiers, which led to more uncertainty about what is likely to follow them.

The other recent study showing a sustained frontal negativity in response to a less predictive context is Kaan and Carlisle's (2014) study comparing responses to letters in highly predictive alphabetic sequences (e.g. *A B C D* ***E***) to letters in non-predictive random sequences (e.g. *C T G W* ***E***). They found more sustained frontal negativity beginning during the P2 time window in response to the second letter in their random sequences, which was the stimulus signalling that the rest of the letters would be unpredictable. Different from both Chou et al.'s (2014) results and ours, though, the sustained part of the response was maximal at the back of the head rather than the front. All three studies have in common that a stimulus that did not provide much constraint about what would follow elicited greater frontal negativity during its P2 time window, and that difference persisted at least until the beginning of the response to the next stimulus. They differ in the scalp distribution of the sustained response, though, which remained frontal in Chou et al.'s (2014) results and ours, while it shifted to a posterior maximum in Kaan and Carlisle's (2014) study. The difference in scalp distribution of the sustained effect could possibly be due to differences in what can be predicted about words and letters in the predictive contexts. The more posterior distribution of the sustained effect in letter sequences could arise from the fact that what can be predicted about meaningless letters is their visual features, which may engage posterior visual processing regions, while what can be predicted about meaningful words is much richer.

Kaan and Carlisle suggested that the sustained frontal negativity they observed in less predictive sequences could be an instance of the SPN, which has been found in situations where one stimulus follows another at a known lag and the first engenders anticipation that the second will provide important information (for a review, see Brunia et al., 2012). The scalp distribution of the SPN varies somewhat with task and type of materials, but it is generally maximal at frontal sites. It has been found to be largest when the first stimulus induces greater uncertainty and the subsequent stimulus is expected to be informative in resolving that uncertainty (Catena et al., 2012), which is consistent with the larger sustained frontal negativity in random letter sequences than in predictive ones in Kaan and Carlisle's study, and in response to less constraining classifiers in Chou et al.'s study and ours.

There is one potential caveat to our interpretation of the response to the classifier, which is related to the non-reliable difference in frequency of occurrence of the general and specific classifiers. High frequency words have been found to elicit smaller P2 amplitudes (Hauk et al., 2006; Hauk & Pulvermuller, 2004; King & Kutas, 1998) so the fact that our general classifiers were more frequent than the specific ones could be one reason they elicited less positivity during the P2 window. However, word frequency effects have not been found to persist the way the difference between general and specific classifiers did here, so frequency differences are an unlikely explanation for the sustained effect of classifier specificity.

To summarise, mismatches between a classifier and the noun it modifies elicited the expected effect on N400 amplitude in English, while the strength of the constraint imposed by the classifier modulated both the size of the N400 effect at the noun and a sustained frontal negativity beginning at the classifier's P2. The latter effect was not specifically predicted, but it fits with other studies finding sustained frontal negativities associated with uncertainty when there is the expectation that the next stimulus will resolve that uncertainty.

The fact that classifier-noun mismatch affected the N400 component and not the P600 component responses to a noun was fully expected for English, given that classifiers are themselves nouns that carry substantial meaning and have many uses other than as classifiers. It is possible, though, that the story would be different for languages like Mandarin because of the ways classifiers differ between languages. Mandarin classifiers are more like bound function morphemes that attach to numbers and quantifiers, rather than standalone words, and they are also used more often in Mandarin because they are obligatory in many more situations than in English. Thus, when a classifier appears in a Mandarin sentence, it is almost certain that it will be followed shortly by a noun that is consistent with it, while in English it may not even be clear yet at a potential classifier whether it is being used as such, making it a much worse predictor of what might follow than in Mandarin. Integrating classifiers and nouns in Mandarin may be more like integrating number or gender between determiners or pronouns and nouns in languages that mark such properties and require them to agree. Violations of agreement typically elicit effects on P600 rather than N400 (Coulson et al., 1998; Friederici et al., 1993; Gunter et al., 2000; Hagoort et al., 1993; Neville et al., 1991), including cases where the basis for agreement is strongly semantic/pragmatic, such as pronouns that are co-referential with an occupation name that is only stereotypically associated with one gender (e.g. *nurse* and *he* or *himself*; Osterhout et al., 1997; Osterhout & Mobley, 1995). Thus, it seems possible that disagreement between classifiers and nouns might elicit P600 effects in Mandarin, either instead of or in addition to N400 effects. It will also be informative about the nature of the sustained frontal negativity elicited by English general classifiers in Experiment 1 to determine whether a similar effect is observed in Mandarin. If this effect is about differences in how much general and specific classifiers constrain what can follow them, the difference might be even larger in Mandarin than in English, in part because what usually follows a classifier in Mandarin is almost always a noun, and in part because some Mandarin specific classifiers are very specific indeed (e.g. *běn – book-like,* which can only be followed by types of books).

Experiment 2

Introduction

A few previous studies of classifier-noun integration have been done in Mandarin and Japanese, which uses classifiers similarly to Mandarin. These studies have yielded inconsistent results about which ERP components are sensitive to classifier-noun integration. In two studies, one in Japanese (Sakai et al., 2006) and one in Mandarin (Chou et al., 2014), people were shown a classifier followed by a noun (not in sentences) and asked to explicitly judge whether they were congruent. In both studies, nouns that were incongruent with the classifiers elicited N400 effects and no P600 effects. In Chou et al.'s study, the degree of constraint provided by the classifiers was also manipulated, using classifiers that impose either weak or strong constraint on what nouns can follow them. They found both a sustained frontal negativity in response to low-constraint classifiers and an interaction between classifier constraint strength and classifier-noun match on the noun's N400, just as we did for English in Experiment 1.

There are reasons, however, to question whether results obtained from word-pair congruity judgments will extend to sentence comprehension. First, it has been demonstrated that task can influence which ERP components are seen in response to words, with congruency judgments maximising N400 effects (Hahne & Friederici, 2002; Oines & Kim, 2014). Second, the same agreement violations that elicit P600 effects in sentences fail to do so when word pairs are used instead (Barber & Carreiras, 2005; Münte et al., 1993). Thus, it is important to investigate classifier-noun processing in sentence contexts using a task that encourages understanding the whole sentence, such as responding to comprehension questions about them. A few studies have done this, but they have additional features that complicate their interpretation, sometimes because of the experimental design and sometimes because of unavoidable properties of Mandarin and Japanese sentences. One Mandarin study (Zhang, Zhang, & Min, 2012) used sentences with a grammatical but highly non-canonical word order with the noun preceding its classifier to allow comparison of responses to the same classifier following different nouns. Zhang and colleagues compared ERP responses to noun-classifier matches (match: *Car he saw a* VEHICLE-CL *black* = *He saw a black car*) with those for two kinds of mismatches, which differed in whether or not they involved a mismatch in animacy (e.g. mismatch with animacy mismatch: *Seal he saw a* VEHICLE-CL *clumsy* = *He saw a clumsy seal* vs. mismatch without animacy mismatch: *Lamp he saw a* VEHICLE-CL *cheap* = *He saw a cheap lamp*). Both kinds of mismatch triggered what was described as an N400 effect (though it was more frontally distributed than usual), which did not differ between the two types of mismatch. However, when analyses were restricted to just the participants who judged the non-canonical word order acceptable, the animacy mismatch condition elicited a P600 effect in addition to the N400 effect.

In a comparison of responses to nouns that violate constraints imposed by classifiers with those that violate constraints imposed by verbs, Zhou et al. (2010) used Mandarin sentences with canonical Subject Verb Object (SVO) word order in which the classified noun was the sentence's direct object and thus followed the verb, so that both verbs and classifiers could precede the critical noun, which could be incongruent with either the verb or the classifier or both. They found an N400 for all mismatch types, but in the conditions that included a classifier mismatch, that difference was followed by a shift to a sustained and more frontally distributed negativity.

Two other studies have investigated classifier-noun processing in Japanese or Mandarin sentences, but before they can be described it is first necessary to explain how relative clauses work in the two languages. First, in both languages relative clauses precede the head nouns they modify (in contrast to English, where relative clauses follow their head nouns). Also in both languages, a classifier can modify an entire complex noun phrase including a relative clause, rather than just modifying the relative's head noun. In such cases, the relative clause intervenes between the classifier and the head noun it modifies. Also in both languages, there is no overt cue marking the onset of a relative clause, as the relative pronouns *who* or *which* do in English. As a result, situations arise where the noun immediately following a classifier is not actually the noun it modifies but is instead part of a relative clause that also modifies the later head noun. This sometimes leads to situations where a noun mismatches the classifier immediately preceding it, but then a later noun appears that is the one the classifier actually modifies. In the Mandarin example below in (4), the classifier *zhī* (for small animals or objects) does not match the noun *mother* following it, because in fact it modifies *apple* later in the sentence.

(4) 一 只 妈妈 买来 的 苹果 …
one zhi-CL mother bought de-MOD apple …
one apple that mother bought …

In Japanese, Mueller, Hahne, Fujii, and Friederici (2005) found both a sustained LAN and a marginal P600 after the onset of a noun that did not match the classifier preceding it. In interpreting Mueller et al.'s (2005) results, Sakai et al. (2006) suggested that the LAN effect may have been due to processing difficulty triggered when a mismatch suggested that there might be a relative clause. They argued that encountering a mismatching noun immediately after a classifier can serve as a cue that a relative clause is coming, thereby increasing processing load and leading to a LAN effect. In Mandarin, Hsu, Tsai, Yang, and Chen (2014) made use of classifiers to investigate relative clause processing, rather than classifier processing *per se*. At nouns that mismatched the immediately preceding classifier, the response was an Anterior Negativity, which the authors interpreted as likely reflecting a combination of the need to resolve the conflict introduced by the mismatch and an increase in processing load caused by the possibility of an upcoming relative clause.

In sum, previous ERP studies of classifier-noun integration in Mandarin and Japanese have found varying effects when nouns mismatch classifiers, including N400, P600, and Anterior Negativities, though N400 effects have predominated. Overall, the pattern of results suggests that processing classifier-noun agreement in Mandarin and Japanese differs from processing morphosyntactic number or gender agreement in languages that have those, but also that it is not just the same as processing a noun following an adjective, since that would be expected to affect just N400 amplitude. In Experiment 2, we investigate classifier-noun integration in Mandarin sentences using the same design and task as in Experiment 1.

Method

Participants

Participants were 33 native speakers of Mandarin (22 female; mean age 21.63, range 18–27) recruited at the University of Illinois. All completed at least their high school education in China and had been living in the USA from 1 month to 6 years, with an average of 12.2 months. They were all strongly right-handed as assessed by the Edinburgh inventory (Oldfield, 1971), had normal or corrected-to-normal vision and no neurological or psychiatric disorders according to self-report. Five had left-handed relatives. All gave written informed consent and received compensation for taking part. Three additional participants (1 female) were run but excluded from analysis due to excessive blinking or motion artefacts.

Materials and design

Forty-two classifiers (21 general, 21 specific) were each used in three different sets of sentences to yield 126 experimental item sets. Each set included three sentence versions: classifier-noun match, classifier-noun mismatch, and classifier missing.[3] The only difference between the match and mismatch versions within a set was the classifier, as illustrated below in Table 2. As in Experiment 1, within each set, if the correct classifier in the match condition was general, so was the incorrect classifier in the mismatch condition, and the same was true for items with specific classifiers. Also as in Experiment 1, the same classifiers and nouns were used in the match and mismatch conditions by re-pairing nouns and classifiers. Stimuli were distributed over three lists such that each classifier was seen twice by each participant, once in the match and once in the mismatch condition, with the two sentences within a list that used the same classifier coming from different item sets. The critical nouns immediately followed the classifier and were identical between sentence versions within a set.[4]

The critical Mandarin sentences in Experiment 2 had a different beginning sequence than the English ones in Experiment 1. The Mandarin sentences began with a locative phrase followed by an obligatory modification particle (*de*) and then the number and classifier, as illustrated in Table 2.

General and specific classifiers were selected based on dictionary entries (Guo, 2002) and specificity was verified with a norming task. Seventy-nine native speakers of Mandarin were asked to produce as many nouns as they could that fit each of the 42 classifiers. The norming task was completed either after the main ERP experiment or after a self-paced reading version of it (whose results are not reported here). Participants came up with more nouns on average for general (mean 3.7, range 1.0–9.4) than for specific classifiers (mean 2.1, range 0.6–6.0; $F(1, 40) = 60.7$, $p < .01$), validating the specificity manipulation.

Critical nouns were matched between items in the general and specific conditions on mean length in number of characters (general: 2.0, specific: 2.0, $F < 1$), mean frequency of occurrence (general: 26.2, specific: 39.9, $F < 1$, measured from the $\text{SUBTLEX}_{\text{CH}}$ corpus; Cai &

Table 2. Sample stimuli.

General	Match	桌子上 /的 /两杯 /咖啡 /已经 /凉了。 On the table/de-MOD/two **cups-CL** /coffee/already /cold. *“The two cups of coffee on the table are already cold.”*
	Mismatch	桌子上 /的 /两张 /咖啡 /已经 /凉了。 On the table/ de-MOD/two **sheets-CL**/coffee/already/cold. *“The two sheets of coffee on the table are already cold.”*
Specific	Match	草地上 /的 /三朵 /野花 /已经 /枯萎了。 In the lawn/de-MOD/three **flower-CL**/flower/already/withered. *“The three flower-like flowers in the lawn have already withered.”*
	Mismatch	草地上 /的 /三阵 /野花 /已经 /枯萎了。 In the lawn/de-MOD/three **wind-CL**/flower/already/withered. *“The three wind-like flowers in the lawn have already withered.”*

Note: Critical words are bolded and underlined. Presentation units are indicated by slashes.

Brysbaert, 2010), and mean number of strokes in the characters (general: 14.9, specific: 15.3, $F < 1$). General and specific classifiers were also matched for length (all composed of one character) and number of strokes (general: 8.2, specific: 8.1, $F < 1$). Just as for the English classifiers in Experiment 1, it was not possible to closely match general and specific classifiers on frequency of occurrence, since general classifiers are used more often because they are used with more different nouns. While the frequency difference was not reliable for the English classifiers in Experiment 1, it was for the Mandarin classifiers in Experiment 2, both for mean overall frequency of the characters (general: 664/million words, specific: 102/million words; $F(1, 124) = 15.0$, $p < .01$) and mean frequency of those characters used as classifiers (general: 172/million words, specific: 31/million words, $F(1, 124) = 45.1$, $p < .01$). (Other uses of some of the characters include noun, verb, adjective, adverb, preposition, etc.) It will again be important to keep the frequency difference in mind when interpreting the ERP responses to the classifiers, given known effects of word frequency on both the P2 and N400 components (Hauk et al., 2006; Hauk & Pulvermuller, 2004; King & Kutas, 1998; Van Petten & Kutas, 1990; Young & Rugg, 1992).[5]

As in Experiment 1, it was necessary to repeat some of the critical nouns in order to use the ones that fit best with each classifier. Nine critical nouns were used more than once, five of them following general classifiers and four following specific classifiers. With one exception, there were at least 20 trials intervening between any two instances of the same noun. The one exception was an instance of a noun repeated after four intervening trials.

Ninety distractor sentences were added to each list for a total of 216 trials/list. There were three types of distractor sentences: (1) Matching classifier + noun sequences occurring at various sentence positions (15; e.g. *A tiger appeared on a mountain-CL mountain*), (2) Mismatching classifier + noun sequences occurring at various sentence positions (15; e.g. *He bought three song-CL houses and made a large fortune*), and (3) Grammatical sentences without classifiers (60; e.g. *There are many cars in the parking lot*). Twenty-five of the distractors began with locative phrases, to try to prevent participants from knowing whether and where to expect classifier-noun sequences to appear in sentences beginning with locative phrases like the experimental items. Each list contained 117 fully correct sentences and 99 sentences with either a classifier-noun mismatch or a missing classifier. Each sentence was followed by a comprehension question that did not specifically probe the comprehension of the classifier. Correct answers to the questions were half *yes* and half *no*. Sentence order was pseudorandomised, with the constraints that there were no more than two critical items in a row and that there were approximately equal numbers of trials in each condition in each of five blocks. Each list was presented in the same order and each participant saw only one list, with equal numbers of trials in each condition (21).

Procedure

The equipment and procedure were identical to Experiment 1, with the following exceptions: (1) stimuli were presented in white 26-point SimSun font at the rate of 450 ms per phrase (350 ms text, 100 ms blank screen); (2) the number and classifier were presented together in a single display because it seemed unnatural to separate them, consistent with the idea that the classifier is a bound function morpheme attached to the number rather than a standalone word; and (3) the obligatory modification particle *de* preceded the number + classifier in Mandarin, whereas *of* intervened between the number and classifier in the English sentences in Experiment 1. The EEG recording session lasted approximately 45 min and the entire session lasted 2–2½ h.

EEG recording and data analysis

All recording and analysis procedures were identical to Experiment 1, with the exception that the epoch extracted from the continuous waveforms for analysis was 1550 ms, from 100 ms before the onset of the number + classifier through 1450 ms later, capturing the responses to the number + classifier and the critical noun. (The epoch extracted in Experiment 1 was longer because *of* intervened between the classifier and the noun.) Trials contaminated with artefacts during this epoch were rejected using the same criteria as in Experiment 1. Epochs contaminated with artefacts were discarded, leading to an average loss of 13% of the data, which did not differ across conditions.

As in Experiment 1, ERP waveforms were analysed using two different baselines: (1) 100 ms before number + classifier onset, and (2) 100 ms before critical noun onset, to allow evaluation of both effects that began at the classifier and continued into the response to the noun and new effects elicited by the noun. The N400 and P600 components were measured using the same time windows as in Experiment 1, but the time window for the P2 component was shifted to 220–270 because P2 peaked at 245 ms in the grand mean waveforms (compared to 255 ms in Experiment 1). The statistical analyses were the same as those used for Experiment 1.

Results

Behavioural results

Comprehension accuracy for questions following target sentences was above 90% for all participants (mean 97%). Response accuracy did not differ reliably depending on classifier specificity (96–97%; *F*(1, 32) = 3.8, *p* > .05) nor on classifier-noun match (96–97%; *F* < 1), nor was there any interaction between match and specificity (*F* < 1).

ERP results

The ERP analyses for this study did not collapse over electrode site laterality as was done for Experiment 1 because there was one reliable interaction with laterality in the results. Thus, the regional analyses are shown at each of six regions in Table 3, rather than just the three shown in Table 1. Figure 3 shows the grand average ERPs for all channels baselined on the 100 ms preceding the number + classifier and continuing through the response to the critical noun and the word following it. Visual inspection revealed that the pattern of responses was similar to Experiment 1, with general classifiers triggering more negative responses, especially at the front of the head, starting in the P2 window and persisting across the epoch, and also with mismatching nouns triggering larger N400 responses. These observations were confirmed by statistical analyses. See Table 3 for ANOVA results.

Number + classifier

At the number + classifier's P2 time window (220–270 ms), the ANOVA over all electrodes revealed a main effect of classifier specificity (*F*(1, 32) = 5.3, *p* < .05) that was modulated by a reliable interaction between specificity and anteriority (*F*(2, 64) = 11.0, *p* < .01) and further modulated by a marginal specificity × anteriority × laterality interaction (*F*(4, 128) = 2.4, *p* < .1). The interactions resulted because items with general classifiers were more negative than those with specific classifiers at frontal sites, and that effect extended back as far as the central sites on the left but not the right (left anterior: *F*(1, 32) = 20.1, *p* < .01; right anterior: *F*(1, 32) = 11.0, *p* < .01; left central: *F*(1, 32) = 15.4, *p* < .01; right central: *F* (1, 32) = 3.8, *p* < .1; left posterior: *F* < 1; right posterior: *F* (1, 32) = 1.1, *p* > .1). The ANOVA on just midline electrodes showed the same interaction between specificity and anteriority (*F*(2, 64) = 5.6, *p* < .05), with general classifiers more negative than specific classifiers at Fz (*F*(1, 32) = 6.0, *p* < .05) and Cz (*F*(1, 32) = 5.6, *p* < .05), but not at Pz (*F* < 1).

The frontal specificity effect that began during the number + classifier's P2 time window persisted throughout its N400 time window, as indicated by a continuing reliable interaction between specificity and anteriority (*F*(2, 64) = 13.3, *p* < .01), as well as a now reliable three-way interaction of specificity, anteriority, and laterality (*F*(4, 128) = 2.8, *p* < .05), in the overall analysis. ROI analyses showed that the specificity effect continued to be fronto-centrally distributed, extending farther back on the left than the right side (left anterior: *F*(1, 32) = 7.4, *p* < .01; right anterior: *F*(1, 32) = 7.1, *p* < .01; left central: *F*(1, 32) = 4.5, *p* = .05).

There were no effects of classifier-noun match in the response to the number + classifier, since the critical noun had not yet appeared.

Critical noun

As in Experiment 1, the three time windows for the critical noun were analysed using two different baselines: (1) 100 ms preceding the number + classifier, illustrated for three midline channels in Figure 4(a), and (2) 100 ms preceding the onset of the noun itself, illustrated for the same channels in Figure 4(b). Both ways of analysing the results are presented to provide an evaluation of both continuing effects beginning at the number + classifier and any new effects triggered by the noun.

Pre-number + classifier baseline. The interaction between specificity and anteriority that began earlier at the number + classifier, with general classifiers more negative at frontal sites, persisted throughout the response to the critical noun. In the noun's P2 window

Table 3. ANOVA *F*-values for Experiment 2.

	Number + classifier (pre-CL baseline)		Noun (pre-CL baseline)			Noun (pre-noun baseline)		
	P2	N400	P2	N400	P600	P2	N400	P600
Overall analysis								
Match (1, 32)	–	–	–	5.8*	3.4!	–	3.0!	–
Spec (1, 32)	5.3*	–	–	3.6!	–	–	–	–
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Spec × Lat (2, 64)	–	–	–	–	–	–	2.7!	4.5*
Match × Ant (2, 64)	–	–	–	–	6.8*	–	–	–
Spec × Ant (2, 64)	11.0**	13.3**	6.2*	9.3**	6.1**	–	–	–
M × A × Lat (4, 128)	–	–	–	–	–	–	–	–
S × A × Lat (4, 128)	2.4!	2.8*	–	–	–	–	–	–
Midline analysis								
Match (1, 32)	–	–	–	5.0*	–	–	–	–
Spec (1, 32)	3.4!	–	–	2.8!	–	–	–	–
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Match × Ant (2, 64)	–	–	–	–	3.6!	–	–	–
Spec × Ant (2, 64)	5.6*	8.4**	3.5!	6.6**	5.5*	–	–	–
Left anterior								
Match (1, 32)	–	–	–	4.2*	6.1*	–	–	6.6*
Spec (1, 32)	20.1**	7.4*	6.3*	9.8**	3.7!	–	–	–
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Right anterior								
Match (1, 32)	–	4.4*	–	5.9*	14.2**	–	–	8.7*
Spec (1, 32)	11.0**	7.1*	–	5.6*	–	–	–	10.0**
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Left central								
Match (1, 32)	–	–	–	5.5*	–	–	4.5*	–
Spec (1, 32)	15.4**	4.5*	5.3*	9.1**	4.3*	–	5.2*	–
Match × Spec(1, 32)	–	–	–	–	–	–	–	–
Right central								
Match (1, 32)	–	–	–	3.6!	–	–	–	–
Spec (1, 32)	3.8!	–	–	–	–	–	–	3.2!
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Left posterior								
Match (1, 32)	–	–	–	3.1!	–	–	6.3*	–
Spec (1, 32)	–	–	–	–	–	–	5.1*	–
Match × Spec (1, 32)	–	–	–	–	–	–	–	–
Right posterior								
Match (1, 32)	–	–	–	–	–	–	–	–
Spec (1, 32)	–	–	–	–	–	–	–	–
Match × Spec (1, 32)	–	–	3.7!	–	–	–	–	–

Notes: Spec = specificity; Ant = anteriority; Lat = laterality; P2 = 220–270 ms; N400 = 350–550 ms; P600 = 600–900 ms.
**$p < .01$.
*$p < .05$.
!$.05 < p < .1$.

the interaction remained reliable in the analysis over all electrodes ($F(2, 64) = 6.2$, $p = .01$), but decreased to marginality at midline electrodes ($F(2, 64) = 3.5$, $p < .1$). In the noun's N400 window, the interaction remained reliable in both the overall ($F(2, 64) = 9.3$, $p < .01$) and midline analyses ($F(2, 64) = 6.6$, $p < .01$), because the specificity effect continued at frontal and central regions (left anterior: $F(1, 32) = 9.8$, $p < .01$; right anterior: $F(1, 32) = 5.6$, $p < .05$; left central: $F(1, 32) = 9.1$, $p < .01$). By the noun's P600 window, the pattern was similar with a reliable interaction in both the overall ($F(2, 64) = 6.1$, $p < .01$) and midline analyses ($F(2, 64) = 5.5$, $p < .05$), but in the ROI analysis it continued to be reliable only at the left central region ($F(1, 32) = 4.3$, $p < .05$).

A main effect of classifier-noun match emerged in the noun's N400 window in both the overall ($F(1, 32) = 5.8$, $p < .05$) and midline analyses ($F(1, 32) = 5.0$, $p < .05$), with the mismatch condition more negative than the match. ROI analyses indicated that the match effect was broadly distributed with fronto-central dominance (left anterior: $F(1, 32) = 4.3$, $p < .05$; right anterior: $F(1, 32) = 5.6$, $p < .05$; left central: $F(1, 32) = 5.5$, $p < .05$; right central: $F(1, 32) = 3.6$, $p < .1$; left posterior: $F(1, 32) = 3.3$, $p < .1$). By the noun's P600 window, the effect of match was reduced to marginality in the overall analysis ($F(1, 32) = 3.4$, $p < .1$), and a match × anteriority interaction emerged (overall: $F(2, 64) = 6.8$, $p < .05$; midline: $F(2, 64) = 3.6$, $p < .1$) because the effect of match remained reliable only at anterior regions (left anterior: $F(1, 32) = 6.1$, $p < .05$; right anterior: $F(1, 32) = 14.2$, $p < .01$).

Pre-noun baseline. To try to separate effects elicited by the critical noun from continuing effects that began at the number + classifier, ANOVAs were also conducted after re-baselining the waveforms on 100

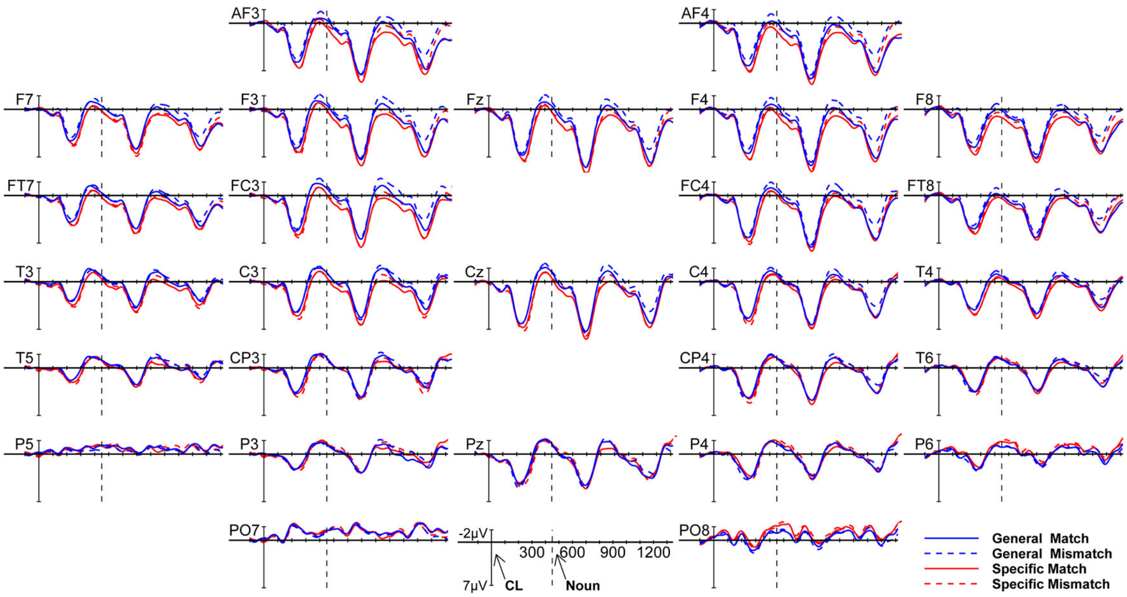

Figure 3. Grand average ERPs at all electrode sites in Experiment 2 baselined on 100 ms before number + classifier. Note: *Y*-axis position indicates onset of number + classifier and dashed line indicates onset of noun.

ms prior to the onset of the critical noun, as shown in Figure 4(b). At the noun's P2 time window, there were no effects of specificity or match (all *F*s < 2.3, all *p*s > .1). Thus, the effects found in this same time window when the waveforms were baselined before the number + classifier appear to be a continuation of effects triggered by the number + classifier. At the noun's N400 window, there was a marginal effect of match in the overall ANOVA (*F*(1, 32) = 3.0, *p* < .1), with mismatch conditions more negative than match conditions. Although there was not a reliable interaction of this effect with anteriority or laterality in the overall ANOVA, it showed the centro-parietal distribution that is typical for N400 (anterior region: *F*(1, 32) = 1.4, *p* > .1; central region: *F*(1, 32) = 3.3, *p* < .1; posterior region: *F*(1, 32) = 4.7, *p* < .05). Although visual inspection

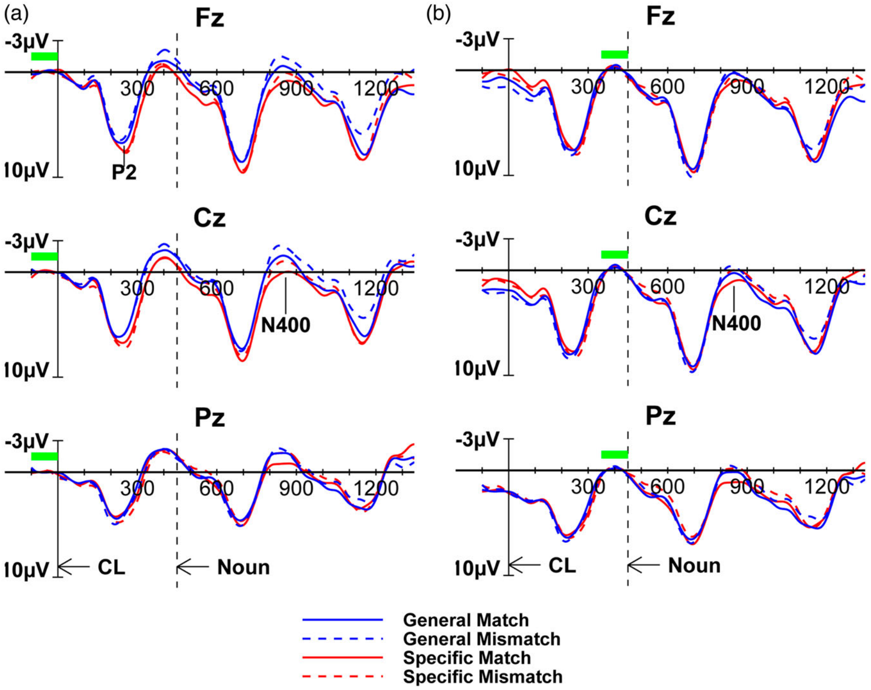

Figure 4. Grand average ERPs at midline sites Fz, Cz, and Pz in Experiment 2 baselined on (a) 100 ms before number + classifier and (b) 100 ms before noun. Note: Noun onset time is indicated by dashed lines and baseline interval is indicated by green bars.

Language, Cognition and Neuroscience 2016.31:761-784.

of the waveforms suggests that the noun's match effect was larger following specific classifiers than following general classifiers, there was no specificity × match interaction in the noun's N400 window ($F < 1$), which is different from the reliable interaction found in Experiment 1. Finally, by the time of the noun's P600 window, the only reliable effect in the overall analysis was a specificity × laterality interaction ($F(2, 64) = 4.5$, $p < .05$) that arose because the noun's waveforms were more negative following specific classifiers than following general classifiers at right frontal sites only ($F(1, 32) = 10.0$, $p < .01$; Fs for all other regions < 1).

To summarise, there was a fronto-centrally distributed specificity effect beginning in the P2 time window of the number + classifier and continuing throughout the epoch, replicating the pattern seen in English sentences in Experiment 1. When the waveforms were baselined on the 100 msec preceding the critical noun, the mismatching nouns elicited a more negative N400 component than the match conditions, also replicating Experiment 1, although the effect was only marginal here. Crucially, there was no evidence of a larger P600 in response to nouns that did not match their classifiers. These results suggest that classifier-noun integration is primarily a semantic process in Mandarin sentences, just as was found for English sentences in Experiment 1.

General discussion

Experiment 2 tested whether classifier-noun integration in Mandarin would differ from English, given differences in the properties of classifiers and the way they are used in the two languages. English classifiers are independent nouns that have many other uses than as classifiers, while Mandarin classifiers are more like bound function morphemes and few of them can be used in other ways. Classifiers are also used far more often in Mandarin because they are obligatory in many more situations than in English. While it seemed quite likely that mismatches between English classifiers and nouns would affect N400 amplitude, the properties of Mandarin classifiers meant it was less clear what to expect for them. One hypothesis was that integrating classifiers and nouns together during Mandarin sentence processing might be more like morphosyntactic agreement processing in languages that mark properties such as number and gender on nouns and various function words that are required to agree with them in sentences. Although classifier-noun integration seems a more semantic process than some of those agreement phenomena in other languages, the fact that violations of gender agreement in English between occupation labels that are stereotypically associated with one gender and co-referring pronouns have been found to affect P600 rather than N400 suggested that P600 might also be sensitive to Mandarin classifier-noun integration. Contrary to that hypothesis, however, our results were very similar for English and Mandarin. In both languages, nouns that did not match the classifier preceding them elicited an increase in N400 and not P600, suggesting that their processing was primarily semantic, and furthermore that classifier-noun integration does not differ substantially from general processes of integrating word meanings while interpreting sentences.

There was one finding in Experiment 1 that was not fully replicated in Experiment 2, which was the interaction between classifier specificity and classifier-noun match, which was reliable at the noun's N400 window in Experiment 1 but not Experiment 2. Whether this is a real difference between the languages is not clear. In both studies, the effect of a mismatching noun on N400 amplitude was numerically larger after more strongly constraining classifiers than more weakly constraining ones, which is most clearly seen at the Pz electrode site in Figures 2 and 4. The size of the N400 mismatch effect was somewhat smaller overall in Experiment 2 (mean 1.5 µV at Cz) than in Experiment 1 (mean 3.4 µV at Cz), which may have precluded detecting an interaction. The mismatch effect was clearly more widely distributed over the head in Experiment 1, but that could just be a consequence of the difference in effect size. There are many factors that could have contributed to this apparent difference in the size of the N400 congruity effect, the first of which is simply that different people were tested in the two studies. Another is that the violations in the mismatching condition could have been somewhat worse overall in English than in Mandarin. It is probably not possible to truly equate the degree of violation even through norming, given differences between the two languages in what classifiers are and how they work. For what it is worth, however, the observed pattern matches the authors' intuitions. That is, the first author finds mismatches equally bad after general and specific classifiers in Mandarin, while the second author finds them worse after specific classifiers in English.

The absence of a reliable interaction between classifier specificity and classifier-noun match in N400 amplitude in the Mandarin study seems at first to be inconsistent with the results reported by Chou et al. (2014) for their Mandarin classifier-noun pairs. However, the interaction in their results came from a condition that we did not include in our study. We included just two types of nouns following the classifiers, which were either one of the most predictable nouns for the classifier or one that did not match the classifier at all. Chou and colleagues also included nouns that were plausible but not predictable for the classifiers, and it was only for those nouns that they found an effect of classifier constraint strength on N400 amplitude. After strongly constraining classifiers, implausible and plausible-but-not-predictable nouns elicited equally large N400s, but after weakly constraining classifiers, there was a graded response with the N400 for the plausible-but-not-predictable nouns intermediate between those for the predictable and implausible nouns. The authors explained the pattern in terms of whether or not some particular noun could be strongly predicted. A strongly constraining classifier affords a specific prediction of a particular upcoming noun (or at least a very small set of highly related nouns), so any noun that is not the predicted one is equally bad and elicits the same size N400. In contrast, a weakly constraining classifier does not afford strong prediction about any particular noun, so whatever noun appears is evaluated for how well it fits with the classifier rather than how well it matches a prediction, leading to a more graded effect.

The other main finding in our results was a sustained frontal negativity elicited by general classifiers compared to specific ones, illustrated for the Fz electrode site in both studies in Figure 5. As a result of this effect, the P2 time window for the critical noun was more positive after the more constraining specific classifiers, but several reasons were raised for not characterising this effect as an instance of the increase in P2 amplitude found in some previous studies for words following more strongly constraining contexts. First, the difference in our critical noun's P2 time window disappeared when the waveforms were baselined before the noun, showing that it was a continuation of an effect that began at the classifier, rather than a new effect triggered by the noun. In contrast, the other studies finding P2 predictability effects at the critical noun found them when the baseline immediately preceded the critical noun itself (Federmeier et al., 2005; Wlotko & Federmeier, 2007, 2011). Second, the effect we found was sustained, beginning during the classifier's P2 time window and persisting throughout the response to the noun, while previous reports of effects of contextual constraint strength on P2 amplitude have been much more temporally restricted (Federmeier et al., 2005; Wlotko & Federmeier, 2007, 2011).

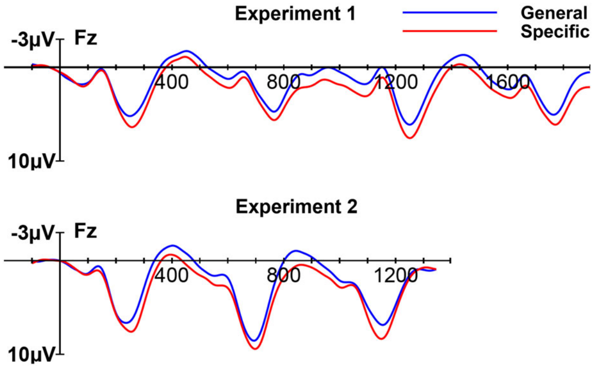

Figure 5. Grand average ERPs at Fz electrode site in General and Specific conditions collapsed across Match and baselined on 100 ms before classifier in Experiments 1 and 2.

A likely reason for the absence of a predictability effect on critical-word P2 amplitude is that our sentences were all presented centrally and such effects have been found almost exclusively when critical words are presented in the Right Visual Field (RVF) and thus processed initially by the left hemisphere. Federmeier (2007) has argued that the reasons such effects only emerge with RVF presentation is that the left hemisphere actively engages in top-down prediction because it has to look ahead in order to accomplish language production, for which it is largely responsible. Right hemisphere processing is argued to be more passively bottom-up, and central presentation yields some undetermined combination of the two modes of processing.

There are some other differences that could be important between our stimuli and procedures and those used in the studies finding effects of contextual constraint strength on critical-word P2 amplitude. One has to do with the source and timing of contextual constraint relative to the critical word. Although our stimuli were sentences, the constraint came entirely from a single word/morpheme, that is, the classifier, which was followed by the critical noun 450 ms later in the Mandarin study and 1000 ms later in the English (because the extra word *of* necessarily intervened between them in English). In contrast, in most of the previous studies finding effects of contextual constraint strength on the amplitude of the critical word's P2 amplitude, constraint was built up across multiple words in sentence contexts, some of them appearing farther in advance of the critical word. Perhaps when the only source of constraint was a single word/morpheme that always immediately preceded the critical noun, there was not time enough to develop the kind of predictions that are reflected in critical-word P2 amplitude modulation. Wlotko and Federmeier (2015) manipulated both the degree of contextual constraint and the presentation rate of words in sentences and found larger effects of constraint strength at 500 ms/word than at 250 ms/word, though it was N400 amplitude and not P2 that was affected by constraint, since the words were presented centrally. Also consistent with the idea that it takes some time to develop predictions, Kaan and Carlisle (2014) found predictability effects on critical-letter P2 amplitude in letter sequences only when an extra 300 ms was added to the usual 500 ms lag between the last context letter and the critical letter. (This study appears to be the only one so far to find a predictability effect on critical-stimulus P2 amplitude using all central stimulus presentation, though the stimuli were single letters rather than words and the constraint effect was sustained and thus may not be the same phenomenon.)

Three other previous studies have examined effects of the degree of contextual constraint supplied by a single-word context on the predictability of a critical word. Huang et al. (2010) and Fruchter et al. (2015) both used single adjectives to provide different kinds/degrees of constraint immediately before critical nouns. Huang et al. used a presentation rate of 1000 ms/word and found effects of adjectival constraint on critical-word P2 amplitude when the critical word was lateralised to the RVF. Fruchter et al. presented their stimuli centrally at a rate of 600 ms/word and found evidence of prediction of the noun in MEG activity during both a 442–600 ms time window after the onset of the adjective but before the noun appeared and a window 197–397 after noun onset. How the latter MEG effect is related to the ERP P2 effects in other studies, though, remains to be determined.

The third previous ERP study to manipulate contextual constraint strength using single-word contexts was Chou et al.'s (2014) study of Mandarin classifier-noun pairs. Their results were strikingly similar to ours in most respects, even though their characterisation of them was somewhat different from ours. They argued that they had found a similar effect of contextual constraint strength on P2 amplitude as previous studies, even though the effect was seen only on the classifier itself and it persisted at least until the noun appeared. (It is impossible to tell whether it persisted across the noun as well, since the waveforms for the noun were plotted using a baseline immediately before it.) They described their results as a P2 effect followed by a sustained frontal negativity, but we have suggested that there is no evidence of two separable effects rather than a single sustained effect, especially given that the scalp distribution of the difference did not change over time. Sustained frontal negativities like this have been found in multiple studies when there is greater uncertainty or ambiguity, such as in non-literal language processing (Coulson & Kutas, 2001) and some kinds of lexical (Federmeier et al., 2007; Lee & Federmeier, 2009; Wlotko & Federmeier, 2012) or referential ambiguity (Nieuwland et al., 2007; Van Berkum et al., 2005). In both Chou et al.'s and our results, a sustained frontal negativity was observed in response to less-constraining general classifiers, which led to more uncertainty about what is likely to follow them.

A possible interpretation of the sustained frontal negativity that we and Chou and colleagues found in response to less constraining classifiers is as an instance of the SPN, as Kaan and Carlisle (2014), suggested for their results for letter sequences. The SPN has been observed in situations where one stimulus follows another at a known lag and the first engenders anticipation that the second will provide important information (for a review, see Brunia et al., 2012). It is largest when the first stimulus induces greater uncertainty and the subsequent stimulus is expected to be informative in resolving that uncertainty (Catena et al., 2012), which is consistent with the larger sustained frontal negativity in response to less constraining classifiers in Chou et al.'s and our studies, since participants knew that another word would follow the classifier after a constant interval and that it would almost certainly be the noun that the classifier modified.

There is a potential caveat to our interpretation of the response to the number + classifier in our results, which is related to the difference in frequency of occurrence of the general and specific classifiers. Recall that it was not possible to closely match the frequencies of the two types of classifiers because general classifiers are inevitably used more often, in part because they are used with many more nouns than are specific classifiers. High frequency words have been found to elicit smaller P2 amplitudes (Hauk et al., 2006; Hauk & Pulvermuller, 2004; King & Kutas, 1998) so the fact that the general classifiers were more frequent than the specific ones could be the reason they elicited less positivity during the P2 window. However, word frequency effects on P2 amplitude have not been shown to persist in the way the difference between general and specific classifiers did here in both studies, so frequency differences are an unlikely explanation for the sustained effects.

There clearly were no effects triggered by mismatching nouns on the P600 component, suggesting that the integration of classifiers and nouns in both English and Mandarin is primarily a semantic process, consistent with Sakai et al.'s (2006) results for Japanese and Zhang et al.'s (2012) and Chou et al.'s (2014) results for Mandarin. The N400 effects observed here were rather small in the Mandarin study, which might seem surprising given that the mismatching nouns were completely wrong for the classifiers they followed. However, the mismatching nouns fit well with everything else in the sentence other than the classifier, since the mismatching sentences were created by replacing the classifier in the match condition with another classifier that was wrong for the noun (but of the same specificity level) in the mismatch condition.

The N400 effect size was considerably smaller than the effects reported by Sakai et al. (2006) in Japanese and Chou et al. (2014) in Mandarin, probably because both of those studies used an explicit congruity judgment task for isolated classifier-noun pairs. Both the task and the absence of any other source of context would be expected to maximise N400 effects.

A final important point to consider is that the Mandarin speakers in this study were immersed in an English-speaking environment and had varying levels of fluency in English. It is possible that they behaved differently than monolingual Mandarins speakers would. It has been demonstrated that parsing preferences in interpreting structural ambiguities in the first language can be affected by acquisition of a second language with conflicting preferences (Dussias, 2004). However, the processing of classifier-noun sequences by Mandarin-English bilinguals seems unlikely to be amenable to such effects, especially given that they are so much more obligatory and frequent in the first language.

In sum, we found that combinatorial processing of classifier-noun sequences in Mandarin and English sentences is primarily semantically based, as indexed by larger N400s when nouns mismatched the classifiers preceding them. We also found in both languages that specific classifiers evoked stronger expectations about what noun might follow them than did general classifiers, as indexed by a larger sustained frontal negativity in response to general classifiers. If our interpretation of the sustained frontal negativity as an index of the degree of uncertainty evoked by the classifier is correct, the same difference between general and specific classifiers should not be observed if it were known already at the classifier what noun it modifies. As it happens, Japanese and Korean allow a test of this prediction because they both have constructions in which the classifier follows its noun rather than preceding it. The prediction is that the same sustained frontal negativity evoked by general classifiers in Mandarin and English will also be found in Japanese and Korean when the classifier precedes the noun, but when the classifier follows the noun, the effect should be absent. Studies in both languages are ongoing in our lab. These studies also allow an examination of whether the N400 mismatch effect that was found at the noun for Mandarin and English also occurs in Japanese and Korean when the mismatch is realised on the classifier because it comes second. It is possible that when the integration of the noun and classifier has to happen on the classifier because it comes second, that could change the integration process in a way that would make the P600 component be the one to be affected by it rather than the N400.

Conclusions

Languages differ in many ways, some of which necessitate different kinds of underlying processes. This may be especially true with respect to what can be predicted, or at least anticipated, at particular points in sentences. Because different ERP components respond to different aspects of language processing, they can provide critical evidence for diagnosing the nature of the various processes underlying comprehension in native speakers of different languages. We have found here that in spite of considerable differences between English and Mandarin in the linguistic properties and obligatoriness of classifiers, and thus in the amount and kind of experience people have had with them, they appear to be processed very similarly in the two languages. We also found in both languages a sustained frontal negativity that seemingly indexes how strongly classifiers constrain what is likely to follow them, which is probably related to similar sustained frontal negativities found in other studies manipulating predictability in sentences and other kinds of sequences. It should be especially useful in future work to compare the predictability of words at different points in sentences between languages that use different word orders, towards coming to better understand the nature of the processes underlying such sustained frontal negativities.

Notes

1. Federmeier et al. (2005) only found this effect when the sentence-final word was presented in the right visual field (RVF), which supports their hypothesis that prediction in language is the province of the left hemisphere because of its primary responsibility for language production, which requires planning ahead (Federmeier, 2007).
2. In most of the studies showing effects of contextual constraint strength on the amplitude of the predictable word's P2, constraint has been built up across multiple words, so there's not a clear single word making the target word predictable (Federmeier et al., 2005; Wlotko & Federmeier, 2007, 2011; cf. Huang et al., 2010).
3. The results are not reported here for the condition with missing classifiers for two reasons. First, because the number+classifier were presented together in a single display in Experiment 2, the response to the one-character number-alone display differed substantially from that to the two-character number+classifier display, in part simply because of the difference in number of characters. Since that display immediately preceded the critical noun, such differences made it impossible to determine an appropriate baseline for comparing responses to the critical nouns in the classifier-missing condition to the other conditions. Second, responses in the classifier-missing condition appeared to change across the

session, so we decided to collect more data to further explore those changes.

4. There were three lists in this study so the match, mismatch, and missing versions of an item were not seen by the same person.
5. We can only assume that the same difference in frequency of occurrence must also have been true for Chou et al.'s (2014) strongly and weakly constraining classifiers, but they did not report the frequency of occurrence of their classifiers.

Acknowledgments

We thank Kara Federmeier, Cynthia Fisher, Jerry Packard, and Darren Tanner for helpful suggestions, and Gabriela Elizalde-Ocala, Michael Hudgins, Young Jae Lee, Raeann Sheley, Keqi Wei, and other undergraduate lab assistants for help with data collection.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by the Dissertation Completion Fellowship and the Network for Neuro-Cultures Graduate Training Fellowship from the University of Illinois at Urbana-Champaign to Zhiying Qian.

References

Aikhenvald, A. Y. (2000). *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi:10.1016/S0010-0277(99)00059-1

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560. doi:10.1080/01690965.2012.658072

Barber, H., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience*, *17*(1), 137–153.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. doi:10.1016/j.brainresrev.2008.05.003

Brunia, C. H. M., van Boxtel, G. J. M., & Boecker, K. B. E. (2012). Negative slow waves as indices of anticipation: The Bereitschaftspotential, the Contingent Negative Variation, and the Stimulus-Preceding Negativity. In S. J. Luck, & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potentials* (pp. 189–207). Oxford: Oxford University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, *5*(6), doi:10.1371/journal.pone.0010729

Catena, A., Perales, J. C., Megias, A., Candido, A., Jara, E., & Maldonado, A. (2012). The brain network of expectancy and uncertainty processing. *PLoS One*, *7*, e40252. doi:10.1371/journal/pone.0040252

Chatrian, G. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *American Journal Electroencephalograph Technology*, *25*, 83–92.

Cheng, L. L. S., & Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry*, *30*(4), 509–542.

Cheng, L. L. S., & Sybesma, R. (2005). Classifiers in four varieties of Chinese. In G. Cinque, & R. Kayne (Eds.), *The oxford handbook of comparative syntax* (pp. 259–292). Oxford: Oxford University Press.

Chierchia, G. (1998). Reference to kinds across language. *Natural Language Semantics*, *6*(4), 339–405.

Chou, C.-J., Huang, H.-W., Lee, C.-L., & Lee, C.-Y. (2014). Effects of semantic constraint and cloze probability on Chinese classifier-noun agreement. *Journal of Neurolinguistics*, *31*, 42–54. doi:10.1016/j.jneuroling.2014.06.003

Chow, W.-Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain Research*, *1506*, 76–93. doi:10.1016/j.brainres.2013.02.016

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2015). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition, and Neuroscience*, doi:10.1080/23273798.2015.106683

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*(1), 21–58. doi:10.1080/016909698386582

Coulson, S., & Kutas, M. (2001). Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters*, *316*, 71–74. doi:10.1016/S0304-3940(01)02387-4

Croft, W. (1994). Semantic universals in classifier systems. *Word*, *45*(2), 145–171.

Dambacher, M., Rolfs, M., Goellner, K., Kliegl, R., & Jacobs, A. M. (2009). Event-related potentials reveal rapid verification of predicted visual input. *PLoS One*, *4*, e50407. doi:10.1371/journal.pone.0005047

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. doi:10.1038/nn1504

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55–64. doi:10.1016/j.bandl.2012.08.004

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, *110*(3), 293–321. doi:10.1016/j.cognition.2008.09.008

Dussias, P. E. (2004). Parsing a first language like a second: The erosion of L1 parsing strategies in Spanish-English bilinguals. *The International Journal of Bilingualism*, *8*(3), 355–371.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. doi:10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. doi:10.1006/jmla.1999.2660

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730–747. doi:10.1016/S0028-3932(01)00193-2

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161. doi:10.1016/j.bandl.2010.07.006

Federmeier, K. D., Mai, H., & Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition*, *33*(5), 871–886.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. doi:10.1016/j.brainres.2006.06.101

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, *1*(3), 183–192.

Fruchter, J., Linzen, T., Westerlund, M., & Marantz, A. (2015). Lexical preactivation in basic linguistic phrases. *Journal of Cognitive Neuroscience*, *27*(10), 1912–1935. doi:10.1162/jocn_a_00822

Gebhardt, L. (2011). Classifiers are functional. *Linguistic Inquiry*, *42*(1), 125–130.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. doi:10.1080/01690960902965951

Greenberg, J. H. (1972). Numeral classifiers and substantive number: Problems in the genesis of a linguistic type. *Working Papers on Language Universals*, *9*, 1–39.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Gullick, M. M., Mitra, P., & Coch, D. (2013). Imaging the truth and the moon: An electrophysiological study of abstract and concrete word processing. *Psychophysiology*, *50*, 431–440. doi:10.1111/psyp.12033

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*(4), 556–568.

Guo, X. (2002). *Xiandai Hanyu Liangci Yongfa Cidian (Dictionary on modern Chinese classifiers)*. Beijing: Yuwen Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483.

Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, *13*(3), 339–356. doi:10.1016/S0926-6410(01)00127-6

Hauk, O., Davis, M. H., Ford, M., Pulvermuller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383–1400. doi:10.1016/j.neuroimage.2005.11.048

Hauk, O., & Pulvermuller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103. doi:10.1016/j.clinph.2003.12.020

Hillyard, S. A., & Münte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Perception & Psychophysics*, *36*(2), 185–198.

Hoecks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. doi:10.1016/j.cogbrainres.2003.10.022

Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(3), 721–742.

Hsu, C.-C., Tsai, S.-H., Yang, C.-L., & Chen, J.-Y. (2014). Processing classifier–noun agreement in a long distance: An ERP study on Mandarin Chinese. *Brain and Language*, *137*, 14–28. doi:10.1016/j.bandl.2014.07.002

Huang, C.-R., & Ahrens, K. (2003). Individuals, kinds and events: Classifier coercion of nouns. *Language Sciences*, *25*(4), 353–373. doi:10.1016/S0388-0001(02)00021-9

Huang, H.-W., Lee, C.-L., & Federmeier, K. D. (2010). Imagine that! ERPs provide evidence for distinct hemispheric contributions to the processing of concrete and abstract concepts. *NeuroImage*, *49*(1), 1116–1123. doi:10.1016/j.neuroimage.2009.07.031

Kaan, E., & Carlisle, E. (2014). ERP indices of stimulus prediction in letter sequences. *Brain Sciences*, *4*(4), 509–531. doi:10.3390/brainsci4040509

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. doi:10.1080/016909600386084

Kaan, E., & Swaab, T. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. doi:10.1162/089892903321107855

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*(2), 113–147. doi:10.1016/j.cognition.2004.01.002

Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. doi:10.1016/S0749-596X(03)00023-8

Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*(1), 37–55.

Kim, A. E., & Gilley, P. (2013). Rapid effects of syntactic prediction during language understanding in visual cortex. *Frontiers in Psychology*, *4*), doi:10.3389/fpsyg.2013.00045

Kim, A. E., & Lai, V. (2012). Rapid interactions between lexical-semantic and word-form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112. doi:10.1162/jocn_a_00148

Kim, A. E., Oines, L. D., & Sikos, L. (2015). Prediction during sentence comprehension is more than a sum of lexical associations: The role of event knowledge. *Language, Cognition, and Neuroscience*, doi:10.1080/23273798.2015.1102950

Kim, A. E., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. doi:10.1016/j.jml.2004.10.002

King, J., & Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience*, *7*, 376–395.

King, J. W., & Kutas, M. (1998). Neural plasticity in the dynamics of human visual word recognition. *Neuroscience Letters*, *244*, 61–64. doi:10.1016/S0304-3940(98)00140-2

Klein, N. M., Carlson, G. N., Li, R., Jaeger, T. F., & Tanenhaus, M. K. (2012). Classifying and massifying incrementally in Chinese language comprehension. In *Count and mass across language (Oxford Studies in Theoretical Linguistics 42)* (pp. 261–282). Oxford: Oxford University Press.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, *5*, 196–214.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*, 1–36. doi:10.1016/S0093-934X(02)00548-5

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 804–823.

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguistic analysis*, *25*(3–4), 209–257.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challengers to syntax. *Brain Research*, *1146*, 23–49. doi:10.1016/j.brainres.2006.12.063

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language? *Language, Cognition, & Neuroscience*, *31*(1), 32–59. doi:10.1080/23273798.2015.1102299

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. doi:10.1016/S0926-6410(03)00086-7

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. doi:10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Lee, C.-L., & Federmeier, K. (2008). To watch, to see, and to differ: An event-related potential study of concreteness effects as a function of word class and lexical ambiguity. *Brain and Language*, *104*(2), 145–158. doi:10.1016/j.bandl.2007.06.002

Lee, C.-L., & Federmeier, K. D. (2009). Wave-ering: An ERP study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. *Journal of Memory and Language*, *61*(4), 538–555. doi:10.1016/j.jml.2009.08.003

Lee, C.-Y., Liu, Y.-N., & Tsai, J.-L. (2012). The time course of contextual effects on visual word recognition. *Frontiers in Psychology*, *3*, 285. doi:10.3389/fpsyg.2012.00285

Lehman, F. K. (1979). Aspects of a formal theory of noun classifiers. *Studies in Language*, *3*, 153–180.

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168. doi:10.1016/j.cortex.2015.02.014

Lewis, A. G., Wang, L., & Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension. *Brain and Language*, *148*, 51–63. doi:10.1016/j.bandl.2015.01.003

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(4), 1–14. doi:10.3389/fnhum.2014.00213

Luck, S. J., & Hillyard, S. A. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, *31*(3), 291–308.

Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, *17*(8), 1229–1244.

Münte, T. F., Heinze, H.-J., & Mangun, G. R. (1993). Dissociation of brain activity related to syntactic and semantic aspects of language. *Journal of Cognitive Neuroscience*, *5*(3), 335–344.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*(2), 151–165.

Nieuwland, M. S., Martin, A. E., & Carreiras, M. (2013). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension. *Brain and Language*, *126*(2), 151–158. doi:10.1016/j.bandl.2013.04.005

Nieuwland, M. S., Otten, M., & Van Berkum, J. J. A. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*(2), 228–236.

Oines, L., & Kim, A. (2014). *Integrate or repair? ERP responses to semantic anomalies depend on choice of processing strategy*. Paper presented at the Architectures and Mechanisms for Language Processing (AMLaP) Conference, Edinburgh.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Language and Cognitive Processes*, *59*(3), 494–522. doi:10.1006/brln.1997.1793

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. doi:10.1016/0749-596X(92)90039-Z

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786–803.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773. doi:10.1006/jmla.1995.1033

Ritchie, W. C. (1971). On the analysis of surface nouns. *Research on Language and Social Interaction*, *4*, 1–16.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, *18*(4), 367–379.

Sakai, Y., Iwata, K., Riera, J., Wan, X., Yokoyama, S., Shimoda, Y., … Koizumi, M. (2006). An ERP study of the integration process between a noun and a numeral classifier: Semantic or morpho-syntactic. *Cognitive Studies*, *13*, 443–454.

Severens, E., Jansma, B. M., & Hartsuiker, R. J. (2008). Morphological influences on the comprehension of subject-verb agreement: An ERP study. *Brain Research*, *1228*, 135–144. doi:10.1016/j.brainres.2008.05.092

Szewczyk, J. M., & Schriefers, H. (2015). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. doi:10.1016/j.jml.2012.12.002

Tanner, D. (2015). On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement. *Cortex*, *66*, 149–155. doi:10.1016/j.cortex.2014.04.007

Tanner, D., Inoue, K., & Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism: Language and Cognition*, *17*(2), 277–293. doi:10.1017/S1366728913000370

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301. doi:10.1016/j.neuropsychologia.2014.02.002

Tecce, J. J., & Cattanach, L. (1993). Contingent negative variation. In E. Niedermeyer & F. Logan da Silva (Eds.), *Electroencephalography: Basic principles, clinical applications, and related fields* (3rd ed., pp. 887–910). Baltimore, MD: Urban & Schwarzenberg.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. doi:10.1037/0278-7393.31.3.443

Van Herten, M., Kolk, H. H. J., & Chwilla, D. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. doi:10.1016/j.cogbrainres.2004.09.002

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380–393.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*, *3*(2), 131–150.

Weckerly, J., & Kutas, M. (1999). An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, *36*(5), 559–570.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. doi:10.1162/0898929041920487

Wlotko, E., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia*, *45*(13), 3001–3014. doi:10.1016/j.neuropsychologia.2007.05.013

Wlotko, E., & Federmeier, K. D. (2011). *Flexible implementation of anticipatory language comprehension mechanisms*. Paper presented at the Cognitive Neuroscience Society Meeting, San Francisco, CA.

Wlotko, E., & Federmeier, K. D. (2012). So that's what you mean! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, *62*(1), 356–366. doi:10.1016/j.neuroimage.2012.04.054

Wlotko, E., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. doi:10.1016/j.cortex.2015.03.014

Wu, Y., & Bodomo, A. (2009). Classifiers≠ determiners. *Linguistic Inquiry*, *40*(3), 487–503.

Young, M. P., & Rugg, M. D. (1992). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology*, *29*(6), 664–676.

Zhang, Q., Guo, C.-Y., Ding, J.-H., & Wang, Z.-Y. (2006). Concreteness effects in the processing of Chinese words. *Brain and Language*, *96*(1), 59–68. doi:10.1016/j.bandl.2005.04.004

Zhang, Y., Zhang, J., & Min, B. (2012). Neural dynamics of animacy processing in language comprehension: ERP evidence from the interpretation of classifier–noun combinations. *Brain and Language*, *120*(3), 321–331. doi:10.1016/j.bandl.2011.10.007

Zhou, X., Jiang, X., Ye, Z., Zhang, Y., Lou, K., & Zhan, W. (2010). Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study. *Neuropsychologia*, *48*(6), 1551–1562. doi:10.1016/j.neuropsychologia.2010.02.001