



Universiteit
Leiden
The Netherlands

Classifiers in Mandarin Chinese: behavioral and electrophysiological evidence regarding their representation and processing

Huang, S.; Schiller, N.O.

Citation

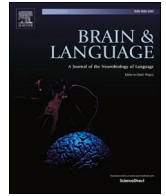
Huang, S., & Schiller, N. O. (2021). Classifiers in Mandarin Chinese: behavioral and electrophysiological evidence regarding their representation and processing. *Brain & Language*, 214. doi:10.1016/j.bandl.2020.104889

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/)

Downloaded from: <https://hdl.handle.net/1887/3142212>

Note: To cite this publication please use the final published version (if applicable).



Classifiers in Mandarin Chinese: Behavioral and electrophysiological evidence regarding their representation and processing

Shaoyun Huang^a, Niels O. Schiller^{a,b,*}

^a Leiden University Centre for Linguistics (LUCL), Leiden University, Reuvenplaats 3, 2311 BE Leiden, the Netherlands

^b Leiden Institute for Brain and Cognition (LIBC), Leiden University, LUMC, Postzone C2-S, P.O. Box 9600, 2300 RC Leiden, the Netherlands

ARTICLE INFO

Keywords:

Chinese classifier feature
Event-related potentials
Overt speech production
Lexical activation
Competitive selection

ABSTRACT

In Chinese, when objects are named with their quantity, a numeral classifier must be inserted between the quantifier and the noun to produce a grammatically correct quantifier + classifier + noun phrase. In this study, we adopted the picture-word interference paradigm to examine participants' naming latencies for multiple objects and their electroencephalogram in four conditions by manipulating two factors, i.e. semantic relatedness and classifier congruency. Results show that in noun phrase production, naming latencies are significantly longer in classifier-incongruent and semantically related conditions than in classifier-congruent and semantically unrelated conditions. Also, an N400-like effect was observed and found to be stronger in classifier-incongruent and semantically unrelated conditions. Together, the behavioral data and event-related potential analyses suggest that the use of classifiers as lexico-syntactic features in Mandarin Chinese takes place via a competitive selection process in noun phrase production.

1. Introduction

The selection process of close-class items has been a topic of controversy in experimental linguistics for decades. Researchers debate over whether a special process takes place that differentiates between the selection of open- and closed-class items (e.g. Dell, 1990; Garrett, 1982; Lapointe & Dell, 1989; Stemberger, 1984; for experimental evidence see Schiller & Costa, 2006; Janssen, Schiller, & Alario, 2014). Studies show that apart from the categorical status of an item, potentially influential elements such as frequency and semantic relationship can also have an impact on the speed of lexical access (see Mahon, Costa, Peterson, Vargas, & Caramazza, 2007). Based on these findings, researchers proceeded further to explore lexical access of close-class items in word-production scenarios. One of the subjects that has received much attention is the determiner, as its selection is often associated with grammatical gender, a lexico-syntactic feature of nouns in numerous languages (see Nickels, Biedermann, Fieder, & Schiller, 2015 for an overview).

1.1. Determiners in Indo-European languages

In speech production, utterances are made up of both open-class

items and close-class items. While open-class items such as nouns and verbs reflect the content the speaker wishes to convey, close-class items mostly serve to construct the utterance in a grammatical way with suffixes, prepositions, determiners, auxiliary verbs, etc. Compared with the selection of open-class items on which a consensus has been reached that semantic information is the key, how close-class items are selected invites more debates (Alario & Caramazza, 2002). Among all the members of this lexical category, the determiner feature in Indo-European languages has been intensively studied during the last few decades. Depending on the language, determiners can be selected according to the properties of the following noun. In English, for instance, the choice of the indefinite determiner is based on the phonological properties of the onset of the following word ("a" for words starting with a consonant and "an" for words starting with a vowel). In languages with more complex morphological structures, the determiner choice involves also factors including grammatical gender and quantity (e.g., in French, "une" and "la" for feminine nouns, "un" and "le" for masculine nouns, and "des" and "les" for plural forms) (see Figs. 1 and 2 and Tables 1–3 in Alario & Caramazza, 2002).

The facilitation or interference effects of those factors on determiner decisions have been tested in numerous studies. Schriefers (1993) left an important note in the history of determiner research as grammatical

* Corresponding author at: Leiden University Centre for Linguistics (LUCL), Leiden University, Reuvenplaats 3, 2311 BE Leiden, the Netherlands.
E-mail address: N.O.Schiller@hum.leidenuniv.nl (N.O. Schiller).

gender congruency has been shown to have an impact on the speed of determiner selection in Dutch, reducing the naming latencies significantly when participants were asked to name pictures presented with distractor words bearing the same gender. However, while Schriefers attributed the increased naming latencies in the incongruent compared to the congruent condition to the competition of grammatical gender between target and distractor, this point was later disputed by researchers arguing that the effect is rather localized in the selection of determiner forms, as it vanishes in the production of plural noun phrases (NP) – a situation where the need to distinguish determiners disappears while the gender feature stays available (Schiller & Caramazza, 2003).

1.2. The gender congruency effect

In early studies on Dutch and German, the gender congruency effect was observed (Friederici & Jacobsen, 1999; Heim, Friederici, Schiller, Rüschemeyer, & Amunts, 2009; La Heij, Mak, Sander, & Willeboordse, 1998; Schiller & Caramazza, 2003, 2006; Schiller & Costa, 2006; Schriefers, 1993; Van Berkum, 1997), but these are languages in which grammatical gender is the only decisive factor for determiner choice in singular nouns. Whether or not the finding can be generalized to other languages where gender is merely one of the cues is another question. This doubt seems to have been confirmed by studies in the Romance language family reporting absence of the effect (Alario & Caramazza, 2002; Costa, Sebastián-Gallés, Miozzo, & Caramazza, 1999; Finocchiaro et al., 2011; for a review see Caramazza, Miozzo, Costa, Schiller, & Alario, 2001), as other features besides gender also have roles to play in the choice of determiner. For instance, Alario and Caramazza (2002) investigated potential gender congruency effect in French using the picture-word interference paradigm (PWI) and found that although participants reacted to the manipulation of phonological resemblance of distractor words by reducing naming latencies in phonologically similar target and distractor scenarios, their performance nevertheless was not affected by grammatical gender congruency. It is therefore proposed that the observed “gender congruency effect” in Dutch and German should actually be interpreted as “determiner congruency effect”.

The same question is also explored in a different design: by assigning bare noun naming tasks, the determiner is now implicit while gender remains an inherent property of the noun. Such tasks allow researchers to investigate the existence of gender congruency effect from another perspective. Findings from experiments within this new spectrum and the above-mentioned NP production studies pointed to opposite directions. In a Dutch bare noun production task, La Heij and his colleagues (1998) failed to find the gender congruency effect, and the results were further attested by the study of Starreveld and La Heij (2004) that reported influence from phonological relatedness rather than gender congruency in the bare noun production experiment. On the contrary, Cubelli’s research team (Cubelli, Lotto, Paolieri, Girelli, & Job, 2005) ran an experiment in Italian and reported that, unlike what was found in studies on Germanic languages, distractor words of the same gender actually displayed an interference effect on target pictures in bare noun naming tasks which disappeared when participants were asked to produce nouns with definite determiners. The inhibitory effect of gender congruency implies that grammatical gender is mandatorily selected in Italian and that the selection was competitive rather than automatic. Such contradictory patterns in Dutch/German and Italian were accounted for by the relative complex morphological structure of the latter (Scalise, 1994; Cubelli et al., 2005), since bare noun naming in Italian involves the inflectional paradigm specified by the final vowel of the word, for example, a feature irrelevant in the production of bare nouns in Dutch. In other words, for bare noun production the selection of grammatical gender goes through a competitive process in morphologically complex languages like Italian but is of minor impact in languages like Dutch (but see also Finocchiaro et al., 2011). For a review on gender congruency effects see Wang and Schiller (2019). A less studied subject is the selection of another lexico-syntactic feature, i.e. classifiers.

1.3. Classifiers and their potential effect in Mandarin Chinese

In Mandarin Chinese, the language under investigation in this study, it is a standard practice to insert a classifier between an article/quantifier and the noun when a noun phrase is produced, and although studies have shown that the choice of classifier is to some extent based on properties of the noun object such as animacy, shape and usage (Shi, 1996; Tai, 1994; Tai & Chao, 1994; Tai & Wang, 1990), there is no single rule derived from conceptual properties of referents that prescribes which classifier to use for a certain category of nouns. In other words, classifiers belong to the class of lexico-syntactic features like grammatical gender. The question of how classifiers are activated and selected during speech production in Mandarin is therefore worth exploring. So far, a limited number of studies has focused on the topic, but research by Wang, Chen, and Schiller (2019) has investigated this issue with both behavioral and electrophysiological means and concluded that the Mandarin classifier feature is automatically activated but not selected in bare noun naming. We thus set up the current study following the same theoretical framework and adopting a similar approach, while making a few changes to the experimental design as well for the purpose of complementing the findings and extending the scope of the previous study.

2. Theoretical framework

2.1. Theoretical accounts of grammatical feature effects

Models of speech production can accommodate the selection of lexico-syntactic features and the gender congruency effect to different degrees. For instance, Levelt, Roelofs, and Meyer (1999) modified the original Word-form Encoding by Activation and VERification (WEAVER) model initially proposed by Roelofs (1992). According to their theory, syntactic properties fall on the intermediate lemma stratum that mediates the conceptual and word-form layers above and beneath it. Given the construction of the three strata, encoding of a word must take place in the order of activation of relevant concepts (semantic category), examination of syntactic and lexical properties (lemma), and retrieval of the word form (phonological representation). A comparable model to Levelt’s model is the Independent Network model proposed by Caramazza (1997), which argues that lemma selection is not in between semantic and phonological information activation, but follows a separate and non-competitive lexical node and that features such as grammatical gender receive activation directly during phonological encoding. Despite the differences between these two renowned models, they nevertheless both break up the process of speech production into a serial procedure. Another line of models, on the other hand, adopts a parallel activation and retrieval view on speech production. The most representative one among these models is the spreading activation model (Dell, 1986, 1988, 1990; Dell & O’Seaghdha, 1991, 1992). Also known as a “connectionist model”, the theory argues that speech production involves the activation of nodes that represent interacting units from concept to phoneme. The interaction of those units does not have a fixed direction and the activation spreads from one node to another until the most activated one is chosen for production.

The explanation for the existence of the determiner effect in Dutch and German NP production studies and lack of significant results in Romance language experiments is yet incomplete without taking into consideration the cross-language differences. Such differences are accommodated by the *late selection hypothesis* (Miozzo & Caramazza, 1999) that offers an account of the presence or absence of the determiner effect. In Germanic languages, grammatical gender alone provides adequate information for determiner choice, hence overt influence on selection time; in Romance languages, however, selection is postponed till the release of phonological properties of the noun since it can sometimes affect the determiner choice as well (e.g., in French: *le chapeau* (masc.) [the hat], *la chambre* (fem.) [the room], *l’oiseau* (masc.)

[the bird], l'étudiante (fem.) [the female student]). Therefore, according to this hypothesis, languages like Dutch and German can be classified as “early selection languages” whereas French and Italian as “late selection languages”.

2.2. The classifier feature in Mandarin Chinese and its comparability to grammatical gender

Unlike the Indo-European language family, Mandarin Chinese does not have grammatical gender in its nominal system; rather, it employs a nominal classifier system that is in some ways comparable to the gender division. In Mandarin Chinese noun phrases, classifiers are preceded by a numeral, a demonstrative or a quantifier and followed by the noun (Li & Thompson, 1989) (e.g., 一把椅子 “yibayizi” [one + classifier + chair]). Classifiers as a nominal feature are similar to grammatical gender of nouns as both are inherent properties of words and neither can be omitted or disregarded for NP production. However, while the assignment of gender to nouns is quite arbitrary and can differ even in closely related languages from the same family, nouns categorized by the same classifier usually do share some perceived similarities. Although no single rule can be imposed on the association of numeral classifier and noun, it is more or less predictable from clues such as animacy, function, shape, size and texture (Allan, 1977; Croft, 1994; Lakoff, 1986; Tai, 1994; Tien, Tzeng, & Hung, 2002; Bi, Yu, Geng, & Alario, 2010). For example, despite being widely different in function, texture, or semantic category in general, 手枪 “shouqiang” [gun] and 牙刷 “yashua” [toothbrush] share the same classifier 把 “ba” due to the fact that they both have a handle. Another distinctive difference between classifiers and grammatical gender is that in Mandarin Chinese there is often an absence of a one-to-one correspondence between nouns and classifiers, whereas in grammatical gender languages this is not as common. For instance, the classifier for 邮票 “youpiao” [stamp] can be 张 “zhang” when it is mainly perceived as “flat and thin”, but another classifier, i.e. 枚 “mei”, can apply as well if the feature “tiny” is highlighted; meanwhile, 棉花 “mianhua” [cotton] as fabric is in most cases associated with the classifier 团 “tuan”, but when the same noun is used to refer to the fruit of the plant, the classifier 朵 “duo” is also legitimate. Finally, classifiers can also change for the same object in different quantities, hence 一条鳄鱼 “yitiaoyu” for “a crocodile” and 一群鳄鱼 “yiquneyu” for “a pack of crocodiles”. The choice of classifiers therefore has a lexico-syntactic nature defined by ambiguous patterns that are flexible to cater for various factors. The major influence of semantic category on classifier assignment has already been attested, with implications that classifiers and nouns undergo similar semantic constraints (Bi et al., 2010; Chen & Wang, 2003), even though the relationship is in many cases opaque (Tzeng, Chen, & Hung, 1991). Nevertheless, even though the use of a classifier preceding a noun depends to a great extent on semantic properties, it is the noun itself in the context that has a dominant effect on the choice of classifiers (Shao, 1993).

Due to the similarities of classifiers and grammatical gender, in recent years researchers have attempted to investigate “classifier congruency effects” upon the same theoretical hypotheses as in determiner/gender congruency-effect studies and compare the results with the empirical findings of the latter. Two nouns are said to be classifier-congruent when they share the same classifier (e.g. 一把扇子, “yibashanzi,” [one fan], 一把刀, “yibadao,” [one knife]). Wang, Guo, Bi, and Shu (2006) examined classifier congruency effects using a PWI paradigm employing both bare noun naming and noun phrase (NP) production tasks. Significant naming latency differences were found when participants were asked to produce NPs, but not bare nouns. This result suggests that the selection of classifiers in Mandarin Chinese does not go through a competitive process in bare noun naming. Moreover, it endorses the morphologically simple/complex distinction made by Cubelli et al. (2005), since Mandarin Chinese has a relatively simple morphological structure. Zhang and Liu (2009), however, conducted a similar experiment and found classifier congruency effects in both bare noun naming and NP production tasks. One of the most recent studies on the

topic is the study by Wang et al. (2019) using a 2 × 2 within-subject design testing whether classifier choice is automatically activated or competitively selected. Analysis of their behavioral data revealed a significant effect on naming latency of bare nouns from semantic relatedness, one of the two factors of the experiment. No effect, however, was found for the other factor, namely the manipulated classifier congruency. The absence of a classifier congruency effect is in accordance with the Wang et al. (2006) study, but since the 2019 study only employed the bare noun naming task, the findings with NP production in former studies were not examined. Wang et al. (2019) also made an advanced step to approach the question with electroencephalography (EEG) analysis. The large negative wave observed in the semantically unrelated condition resembled an N400 effect and resonated with their behavioral data findings. More importantly, a significantly stronger N400 effect was also observed in the classifier-incongruent relative to the classifier-congruent condition, which Wang et al. (2019) interpreted as a result of the spreading activation from activated lemmas. In short, the study concluded that the classifier feature in Mandarin Chinese is automatically activated, but not selected.

2.3. The current study

Following the theoretical framework laid out in the Wang et al. (2019) study, we collected behavioral data (naming latencies) and ERP data while manipulating two factors, i.e. semantic relatedness and classifier congruency, in a PWI paradigm. We will therefore make predictions for our results based on the empirical evidence regarding the two factors, respectively.

It is generally accepted in speech production research that the production of a word in the presence of distractors involves a process of competitive selection. Known as *lexical selection by competition*, this hypothesis proposes that the more levels of non-target words are activated, the longer it takes to select the target word (e.g., Belke, Meyer, & Damian, 2005; Bloem & La Heij, 2003; La Heij, 1988; Levelt et al., 1999; Roelofs, 2003). When a participant is presented with a semantically related distractor word, activation of lexical nodes occurs not only from its written form, but also via the target picture that is conceptually similar, hence more levels of activation of the distractor word and longer latencies to name the target picture. According to this theory, we expect to find shorter naming latencies in the semantically unrelated condition than in the semantically related condition, a prediction that is also in accordance with the results from the Wang et al. (2019) study.

With regard to the employment of the ERPs analysis, the N400 effect is a negative component peaking around 400 ms post-stimulus with centro-parietal maximum that is usually taken to indicate semantic anomaly or difficulty to integrate words in the discourse, semantic unrelatedness, as well as word frequency (e.g., Ganushchak, Christoffels, & Schiller, 2011; Glaser & Dünghoff, 1984; Koester & Schiller, 2008; Kutas & Federmeier, 2011; Leckey & Federmeier, 2019; Rugg, 1990; Strijkers, Holcomb, & Costa, 2011). Therefore, semantic unrelatedness should evoke a larger negative wave in the time window around 400 ms post-stimulus that is particularly robust in centro-parietal regions. Nonetheless, we should bear in mind that the N400 effect is quite sensitive to repetition (Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991), even though it is common practice to repeat targets and distractors in a PWI paradigm (for a review, see Glaser, 1992), and this is generally preferred over a between-subjects design where each participant names a particular target in one condition only. Since each of the target pictures in the current study appears eight times and each distractor word twice, we may possibly extract the negative wave with a reduced amplitude.

The most prominent difference between the current study and the Wang et al. (2019) study lies in the nature of tasks, i.e. the current study required participants to name pictures through NP production whereas the Wang et al. (2019) study deals with simple bare noun naming. In other words, while classifiers for both target pictures and distractor

words were not explicitly produced in the Wang et al. (2019) study, our task required participants to explicitly produce the correct numeral classifiers for the target pictures. Consequently, and contrary to the Wang et al. (2019) study, we can rule out the possibility of non-activation of classifiers for target concepts. The two possibilities left are automatic activation and competitive selection.

If the former turns out to be the case, differences in naming latencies should not be observed. Moreover, the same N400 effect that reflects semantic integration difficulty has been proven to be a robust effect associated with classifier and noun inconsistency, no matter whether the NP appears in sentences or in isolation, and whatever its position, order and classifier-noun distance is (see Tsai, Hsu, Yang, & Chen, 2008; Zhou et al., 2010; Zhang, Zhang, & Min, 2012; Hsu, Tsai, Yang, & Chen, 2014). Since the Wang et al. (2019) study found a stronger N400 effect in the classifier-incongruent condition compared to the congruent condition, presumably due to the automatic activation of classifiers, in the present study we expect to also see an N400-like effect within a similar time window. The predicted electrophysiological observation under this possibility is supported by research from Barber and Carreiras (2005) who found similar effects of gender incongruity for covert production of noun phrases. In short, the lack of differences in naming latencies and the ERP effect around 400 ms would suggest automatic activation in NP production. However, it is important to note that although the automatic activation of classifiers in NP production may not be unusual, null effects of congruency in both NP and bare noun production will make Mandarin Chinese a unique language that is different from both the morphologically more complex Romance languages and the relatively more simple Germanic ones. Moreover, it will also contradict the findings from the Wang et al. (2006) and Zhang and Liu (2009) studies.

In the alternative case, if classifier access is a rather selective process in NP production in Mandarin Chinese, the dichotomy of classifier congruency for target pictures and corresponding distractor words would lead to significant differences in naming latencies. Specifically, naming latencies should be shorter in the classifier-congruent condition than in the incongruent condition. Likewise, we would also predict an N400-type effect that is stronger in classifier-incongruent trials than congruent ones. This together with findings from the previous study would support Wang et al. (2006) conclusion that classifier encoding in Mandarin Chinese is only necessary in NP production. Furthermore, the selection and bypassing of classifiers in NP production and bare noun naming, respectively, suggest that Mandarin Chinese resembles German and Dutch in terms of the encoding of lexico-syntactic features in speech production.

One thing that was not deeply investigated in the relevant studies in Mandarin Chinese speech production, however, are the implications of the existence of interactions (or lack thereof) between semantic relatedness and classifier congruency. The feed-forward, serial account of speech production represented by the WEAVER and Levelt's model (Roelofs, 1992, 1993, 1997, 2003; Levelt et al., 1999; Levelt & Schriefers, 1987) proposes that lexical selection and determiner choice takes place in an irreversible sequential order. Since the classifier feature, like grammatical gender, is dominated by individual nouns (Shao, 1993), its selection or activation cannot precede the processing of nouns. Semantic relatedness would affect the speed of noun selection and thereby determiner selection because lemma selection occurs in between semantic and phonological information retrieval. As a result, a significant effect from semantic relatedness on naming latency is expected, but this effect is independent of the classifier congruency itself. Another representative serial model, the Independent Network model (Caramazza & Miozzo, 1997), would suggest no interaction between the two factors either, given that lemma selection follows a separate lexical node and syntactic features like grammatical gender receive activation from phonological activation. However, if speech production, as argued by parallel-processing models, is a cascading process that allows interactions and convergence in any direction in between nodes and levels of representation (see Dhooge, De Baene, & Hartsuiker, 2016 for a summary), activation of a lexical unit would activate the corresponding

classifier too, and the presence of distractor words would cascade activation to their respective classifier nodes. When the target picture and distractor word belong to the same semantic category, the distractor has more sources of activation, hence more activation of its classifier as well. If classifier choice is competitive, the classifier-congruent condition would witness a facilitation effect on the target classifier from the distractor. In the classifier-incongruent condition though, extra activation of the non-target classifier would lead to more competition and slow down the selection process. On the other hand, when the target picture and the distractor word are semantically unrelated, both have lower levels of activation. Consequently, in the classifier-congruent condition there may still be spreading activation, but not as much as in the case of semantic relatedness. In the classifier-incongruent condition, there would be a smaller delaying effect as a result of less activation of the distractor.

3. Methods

3.1. Participants

Twenty-five native Mandarin Chinese speakers (mean age = 24, SD = 2.9; 21 females) studying at Leiden University in the Netherlands as registered students or exchange students were recruited for the experiment. All of them were native Mandarin speakers who also speak one or more other languages. Notably, some of them were born and raised in southern regions of China where local dialects are spoken at home¹. Participants were given informed consent forms to read and sign before participating in the experiment. All of them had normal or corrected-to-normal vision and none reported a history of brain impairments or surgeries. Participants were paid for their participation, regardless of the level of performance.

3.2. Materials

Thirty objects were carefully selected to ensure that each referred to a familiar and concrete concept. Black-and-white line drawings were used to represent the objects. Twenty-seven percent of the intended names of the pictures were monosyllabic words, three percent were trisyllabic and the rest were all bi-syllabic. Some of the target objects or their corresponding pictures were replaced with those that participants could more easily agree upon, according to observations from pilot studies.

Four distractor words were chosen for each target, resulting in 120 combinations of target and distractor pairs in total. Distractors differed in whether or not they belonged to the same semantic category and whether or not they were associated with the same classifier as the targets. Distractors were strictly controlled for word frequency based on the Modern Chinese Frequency Dictionary (1986) and visual complexity determined by number of strokes (average strokes of distractor words are restricted to a narrow range from 14.7 to 16.6 across the four conditions). Distractors were not phonologically or orthographically related to each other or to the targets (for discussions around phonetic and/or visual complexity, see Lupker, 1982; Qian, Reinking, & Yang, 1994).

The categorization for semantic relatedness between targets and distractors was validated. A total of nine Chinese Mandarin native speakers who had not further been involved in the experiment were asked to rate the semantic relatedness of target – distractor pairs on a scale of 1–10. The average scores for classifier-congruent and classifier-incongruent conditions were 5.51 and 5.07, respectively, showing that the semantic categories of stimuli were not associated with congruency. Meanwhile, the average scores for semantically related and unrelated

¹ Most of the participants were from the eastern coastal major cities where Mandarin education was reinforced. Participants whose performance showed obvious influences from dialects were not included in the analysis.

pairs were 8.31 and 2.27. With a Wilcoxon signed-rank test, the difference between the ratings of the semantic relatedness for classifier-congruent and incongruent pairs was shown to be statistically insignificant ($p = 0.06$), but for the semantically related and unrelated pairs the difference was significant ($p < 0.001$). In other words, the semantic relatedness of target-distractor pairs between the two semantic conditions was indeed perceived differently, but the two classifier conditions were not perceived to differ in terms of semantic relatedness.

The classifiers were checked for their meanings and conditions of usage in the online app of *Xinhua Zidian* (11th edition) in order to be selected as candidates. The combination of nouns and classifiers was then examined in three steps as to make sure that the choice of classifier – noun pairs raise the least controversies. First, the pairs were reviewed by one native speaker from Shandong province, China, who majored in Chinese linguistics. Second, pilot studies were run to replace the less agreed pairs with those for which more consensus could be achieved. Finally, after the familiarization session of the experiment, participants were provided with the suggested classifiers for the nouns to which they had assigned different classifiers themselves. They would be asked whether they felt the pairs sounded natural, and if they did not think so, they were allowed to use their preferred classifiers and we removed the stimuli from their data in the analysis. In most cases, though, participants agreed that our pairs were more or equally natural and stuck to the combinations in the experimental session.

3.3. Design and procedure

The experiment had a 2 by 2 within-subject design. Classifier congruency (C) and semantic relatedness (S) were the two main factors with two levels each, i.e. congruent (+) or incongruent (–), related (+) or unrelated (–), resulting in four conditions for the target and distractor combination: C+S+, C+S–, C–S+ and C–S– (see Table 1).

In each trial, either two or three identical pictures appear simultaneously on the screen. In other words, participants saw target pictures presented in pairs of two or three in each condition, resulting in eight appearances of each target. Each participant therefore went through 240 recorded trials in total.

The Windows program Mix (Van Casteren & Davis, 2006) was used to assign a pseudo-random order to the trials. After each trial, a minimal distance of ten other trials was inserted until the same target appeared again. No two targets associated with the same classifier or two identical conditions with different targets were allowed to appear in consecutive trials. Trials with the same number of pictures were made to appear at most twice in a row in order to prevent priming effects from preceding trials. The pseudo-randomization procedure was repeated once the same trial order has been used twice, so that potential ordering effects could be counterbalanced.

The complete experiment consisted of three sessions. Participants began with the familiarization session in which they were taught the intended names of the target items by slides containing a picture of the item and its name underneath. Each slide stayed on the screen for 3000 ms. After all 30 slides had been presented, participants were informed of

the upcoming practice session by a slide announcing that they should now expect to see two or three identical pictures at the same time without the names. In the middle of the screen, the meaningless letter string “XX” would appear surrounded by the pictures. Participants were asked to ignore the string “XX” and name the pictures with the name they had learnt in the familiarization session. An example was given before the 30 slides to make sure that participants knew the expected way of naming, i.e. “number-classifier-item”. They were corrected by the end of the practice session if they used wrong names or unintended classifiers.

After the practice session, participants started the experimental session. The experimenter explained, along with instructions on screen, that participants were expected to name the pictures, just as they did in the practice session, regardless of the distractor words. The 240 experimental trials that followed were divided evenly into four sessions, with a break in between sessions (the length of which was determined by participants). Each trial started with the presentation of a fixation cross “+” for 300 ms and then switched to a blank screen for another 300 ms. Then the slide with target pictures and distractors was presented for 3000 ms and faded out no matter whether or not a response was given. The slide was followed by another display of a blank screen for 500 ms as the closure of the trial. At the beginning of the experiment as well as after each intermission, a “warm-up” trial with non-target pictures was presented, without informing the participants that their responses to these trials would not be recorded.

Participants were placed in front of a computer screen in a sound-proof recording booth that was dimly lit. At the beginning of the practice session as well as each of the four sessions in the experimental part, texts were displayed on the screen to emphasize that participants should respond in a fast and accurate manner. Audio recordings were made for the entire period of the 3000 ms slide of each trial for measurement of response time using Praat 6.0.49v (Boersma & Weenink, 2019) later. The electroencephalogram (EEG) was also recorded during the experimental session of the experiment.

3.4. Electrophysiological recording and data processing

Thirty-two Ag/AgCl electrodes on the standard scalp sites of the extended international 10/20 system were used to measure and record EEG data. A total of six flat electrodes were attached to each participant’s facial skin to record electrical potentials. Four of them were placed around the eyes to detect blinks and horizontal eye movements and two at the mastoid positions to be used for the purpose of re-referencing the acquired data.

Brain Vision Analyzer 2.0.4v (Brain Products GmbH, 2013) was chosen as the offline processing tool for the EEG data. The signals were first re-referenced offline to the average of the left and right mastoid, with a high- and low-pass filter that cut off the band at 0.1 Hz and 30 Hz (24 dB/cot). Sampling rate of all datasets was down-sampled from 512 Hz to 256 Hz to enable better comparison with the analysis of the Wang et al. (2019) study. Semi-automatic Ocular Correction ICA was performed to detect and fix artifacts due to blinks. Through data inspection

Table 1

Example of slides used in the experimental session in all the classifier congruency (C) and semantic relatedness (S) conditions.





Target picture KNIFE	Condition			
	C+S+	C–S+	C+S–	C–S–
Classifier /ba3/ 把	 叉子	 盘子	 扇子	 雪茄
Distractor	<i>fork</i> /cha1zi0/ 叉子	<i>plate</i> /pan2zi0/ 盘子	<i>fan</i> /shan4zi0/ 扇子	<i>cigar</i> /xue3jia1/ 雪茄
Classifier of distractor	/ba3/ 把	/ge4/ 个	/ba3/ 把	/zhi1/ 支

Table 2

General mixed effects model results of naming latency with classifier congruency and semantic relatedness as two predictors. Random effects from both participants and items were considered. Results of analysis using the lme4 package in R reveal that both classifier congruency and semantic relatedness had a significant influence on naming latencies.

Predictors	t Resp			t Resp			t Resp		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	0.87	0.82–0.93	<0.001	0.87	0.81–0.92	<0.001	0.87	0.82–0.93	<0.001
Incongruent				0.01	0.00–0.02	0.040	0.01	0.00–0.02	0.041
Unrelated							–0.01	–0.02 to –0.00	0.026
Random Effects									
σ^2	0.03			0.03			0.03		
τ_{00}	0.00 _{Item}			0.00 _{Item}			0.00 _{Item}		
	0.01 _{Subject}			0.01 _{Subject}			0.01 _{Subject}		
ICC	0.37			0.37			0.37		
N	21 _{Subject}			21 _{Subject}			21 _{Subject}		
	30 _{Item}			30 _{Item}			30 _{Item}		
Observations	4701			4701			4701		
Marginal R ² /Conditional R ²	0.000 / 0.374			0.001 / 0.375			0.001 / 0.376		

and artifact rejection we also made sure to exclude trials including voltage steps larger than 50 μ V and the absolute value of amplitudes exceeded 100 μ V within a single interval of 200 ms. Because the N400 effect is often most prominent in centro-parietal regions, only relevant channels in left and right fronto-central and left and right centro-parietal regions (F3, FC1, FC5, C3, CP1, CP5, P3, PO3, F4, FC2, FC6, C4, CP2, CP6, P4, and PO4) were enabled for data analysis. We further segmented the data into four groups corresponding to the four conditions using pre-assigned triggers embedded in the experiment design. Epochs from –200 ms pre-stimulus onset to 700 ms post-stimulus onset were extracted, with baseline correction using the –200 ms to 0 ms interval. We had to discard data from some participants due to excessive artifacts and electrode drift due to poor connection, bad signals, abnormal amplitudes and too much head movement. Notably, some of these participants had an acceptable artifact rate, but when trials containing incorrect responses were also excluded, the number of remaining usable epochs were considered insufficient with more than 40% data loss or unbalanced available trials across conditions. In the end, a total of 15 clean datasets were left for ERP data analysis.

4. Results

4.1. Behavioral data

Four participants were excluded from analysis as a result of poor performance, therefore the data of twenty-one participants were analyzed (5,040 trials). Of all recorded data entries 6.71% were dropped from the behavioral data analysis because of (a) participants giving incorrect responses or stuttering (5.40%); (b) participants giving correct responses in only one or two out of all eight cases of a target, showing an established association of an unintended classifier with the target item (0.10%); (c) outliers (i.e., missing responses or seriously delayed responses due to unexpected external distractions; naming latencies exceeding 3 SDs around the participant's average response time; 1.21%).

Given that both our participants and items only represented a random sample, we must control for the randomness and avoid the risk of increasing likelihood of a Type I error due to the unaccounted variability of participant and item choice (for a detailed discussion, see Appendix B). Instead of the classic repeated measures ANOVAs test that only submits condition means across either participants or items to analysis, the general mixed effects regression was employed in this study to control for potential random effects of both participants and items simultaneously.² The analysis was run in R (R Core Team, 2012) along

² Nevertheless, ANOVA analyses were still performed (see Appendix B) for the sake of comparison with the Wang et al. (2019) study.

with the “lme4” package (Bates, Maechler, & Bolker, 2012) with response time (RT) as a function of classifier congruency (same classifier vs. different classifiers) and semantic relatedness (same semantic category vs. different categories).

The model (see Table 2) demonstrated significant results with classifier congruency and semantic relatedness as fixed effects and intercepts for participants and target objects as random effects. A main effect of classifier congruency was found on RTs ($\chi^2(1) = 4.189$, $p = 0.041$), reducing the naming latency by 9.9 ± 4.8 ms when the target and the distractor have the same classifier. Semantic relatedness also had a main effect on RTs ($\chi^2(1) = 4.981$, $p = 0.026$), prolonging the naming latency by 10.8 ± 4.8 ms when the target and the distractor belong to the same semantic category. Possible interaction of the two main factors has also been examined, but no significant effect was found on RTs ($\chi^2(1) = 1.042$, $p = 0.307$) (see Fig. 1).

4.2. ERP data

Of all experimental trials 26.44% were discarded from the analysis due to incorrect responses (6.16%) and artifact rejection (20.28%). For each condition, an average of 49 epochs was available for analysis. The data were categorized into four regions of interest (ROIs): left fronto-central (F3, C3, FC1, FC5), right fronto-central (F4, C4, FC2, FC6), left centro-parietal (CP1, CP5, P3, PO3) and right centro-parietal (CP2, CP6, P4, PO4). This categorization is kept the same as in Wang et al. (2019) previous report for the purpose of comparison. Epochs in each trial were divided into three consecutive windows (0–170 ms, 170–385 ms and 385–585 ms) based on visual inspection of ERP peaks and previous studies (Costa, Strijkers, Martin, & Thierry, 2009; Dell'Acqua et al., 2010; Zhu, Damian, & Zhang, 2015) in order to avoid biasing ERP component measurement procedures toward significant but bogus effects (Luck & Gaspelin, 2017). Repeated analysis of variance (ANOVA) was conducted separately for every time window on mean amplitudes of segmentations for all selected channels with the R software (R Core Team, 2012) using the ez package (Lawrence & Lawrence, 2016). The two independent variables of classifier congruency (two levels) and semantic relatedness (two levels) that were used in the behavioral data analysis were also used as predictors in the ERPs analysis, together with ROIs (four levels).

Repeated ANOVA (see Table 3) were run in time windows 0–170 ms, 170–385 ms and 385–585 ms, respectively. Since the mean amplitudes obtained were averaged across items, only the by-subject analysis was performed. In the first two time windows, neither classifier congruency nor semantic relatedness yielded significant effects (although in the second window an interaction between them was observed), but ROI did have an influence on the amplitudes, reflecting the fact that the fronto-

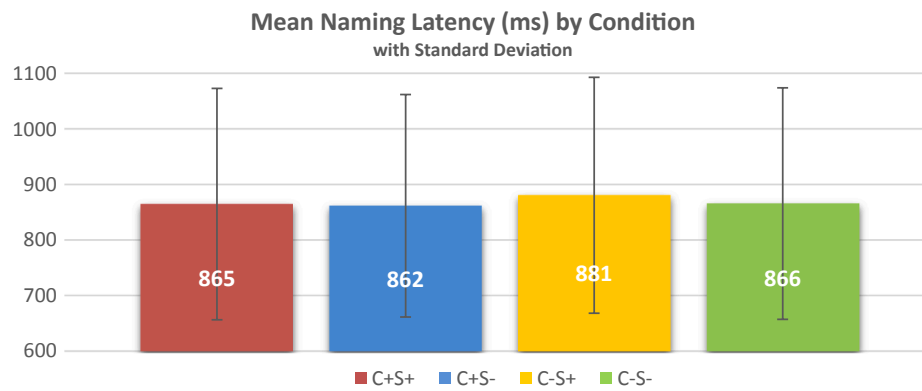


Fig. 1. Naming latency averaged across the semantic relatedness the classifier congruency conditions. Naming latencies in the semantically related condition were significantly longer than those in the semantically unrelated condition. Similarly, naming latencies in the classifier-incongruent condition were significantly longer than those in the classifier-congruent condition. There was no interaction between semantic relatedness and classifier congruency.

central areas bilaterally displayed enhanced negativity compared to the centro-parietal areas bilaterally. In the time window 385–585 ms, we observed a negative wave that peaked approximately between 420 and 460 ms, and a main effect was found from classifier congruency, $F(1, 14) = 4.96$, $p = 0.043$, semantic relatedness, $F(1, 14) = 9.64$, $p = 0.008$, as well as ROI, $F(3, 42) = 5.04$, $p = 0.005$. This denotes an N400-like effect that was stronger when target and distractor classifiers were incongruent than when the classifiers were congruent. Similarly, the

Table 3

(a–c) Results for repeated ANOVA of mean amplitudes of epochs in time windows 0–170 ms, 170–385 ms and 385–585 ms, with classifier congruency, semantic relatedness and ROIs and their interactions as predictors.

(3a) By-subject ANOVA in time window 0–170 ms			
Source	F		
	DFn	DFd	F
Classifier Congruency	1	14	0.12
Semantic Relatedness	1	14	0.27
ROI	3	42	11.13*
Congruency × Relatedness	1	14	0.42
Congruency × ROI	3	42	0.90
Relatedness × ROI	3	42	0.24
Congruency × Relatedness × ROI	3	42	0.16
(3b) By-subject ANOVA in time window 170–385 ms			
Source	F		
	DFn	DFd	F
Classifier Congruency	1	14	0.01
Semantic Relatedness	1	14	1.03
ROI	3	42	11.14*
Congruency × Relatedness	1	14	5.23*
Congruency × ROI	3	42	0.74
Relatedness × ROI	3	42	0.49
Congruency × Relatedness × ROI	3	42	0.17
(3c) By-subject ANOVA in time window 385–585 ms			
Source	F		
	DFn	DFd	F
Classifier Congruency	1	14	4.96*
Semantic Relatedness	1	14	9.64*
ROI	3	42	5.04*
Congruency × Relatedness	1	14	4.19
Congruency × ROI	3	42	1.49
Relatedness × ROI	3	42	0.24
Congruency × Relatedness × ROI	3	42	0.67

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

result also shows that there was a more negative wave within the classical N400 window when target and distractor stimuli were semantically unrelated than when they were from the same semantic category. No two-way or three-way interactions between any of the three factors were found in this time window after sphericity corrections were applied (see Fig. 2).

5. Discussion

In this study, the potential impact of classifier congruency and semantic relatedness on speech production in Mandarin Chinese was examined by employing the PWI paradigm. As participants saw both the target pictures and corresponding distractor words simultaneously on the screen, we recorded their picture naming latencies and real-time EEG activities. We will discuss the activation and selection of target nouns based on the behavioral and ERP data we have attained, starting with the semantic effect.

Our behavioral results are in accordance with the established findings that distractor words from the same semantic category as the target picture cause longer naming latencies through imposing a semantic interference effect (e.g. Glaser, 1992; La Heij, 1988). Similar results were obtained in the previous study by Wang et al. (2019) in both by-item and by-subject analyses using the repeated ANOVA measures. These all point to the conclusion that semantic relatedness results in more competition during the lexical selection process, as more than one candidate is activated at the same time, resonating with the lexical selection-by-competition view shared by numerous studies that found semantic interference (e.g., Belke et al., 2005; Bloem & La Heij, 2003; La Heij, 1988; Levelt et al., 1999; Roelofs, 2003).

On the other hand, our ERP data on fronto-central and centro-parietal regions display a negative wave within 385–585 ms post-stimulus onset. This time window largely overlaps with the classic N400-effect window, and the negativity is believed to be evoked by the presence of distractor words while the target picture was displayed. Crucially, the wave is significantly more negative in the semantically unrelated condition than in the semantically related condition. The outcome is in line with the Wang et al. (2019) study which also observed a significantly more negative ERP wave reflecting an N400 effect in the semantically unrelated condition, as well as with other PWI-based ERPs studies focusing on the semantic interference effect in Mandarin Chinese and other languages (e.g., Dell'Acqua et al., 2010; Wicha, Moreno, & Kutas, 2003; Zhu et al., 2015).

Meanwhile, we also manipulated classifier congruency to test whether or not it has an influence on accessing the target word in the presence of a distractor. Results from our linear mixed effects model suggest that participants were significantly faster to name the target

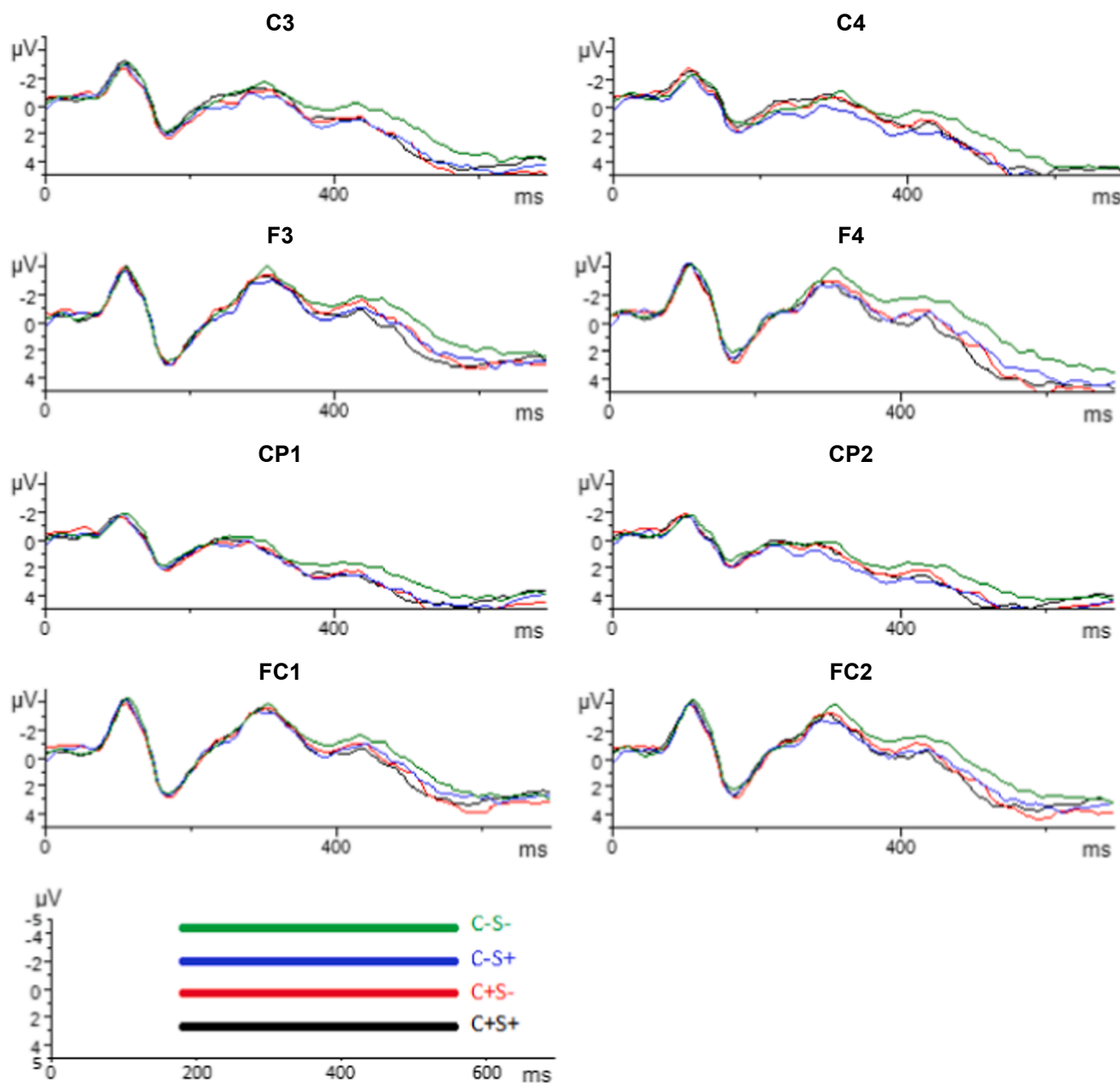


Fig. 2. Grand averages of epochs from representative electrodes for the comparison across the four conditions.

picture in the classifier-congruent than in the classifier-incongruent condition. Notably, no significant difference was found in the previous study, but the discrepancy is supported by the research of Wang et al. (2006) who found significant classifier congruency effects only in NP production tasks. The fact that a significant difference in naming latencies was absent in the Wang et al. (2019) study but found in the current one may have been caused by several reasons.

First of all, in the previous study by Wang et al. (2019), participants were required to carry out a bare noun naming task, and according to the study's conclusion, the classifier feature of Mandarin Chinese words is automatically activated, but not selected during bare noun naming (because classifiers are not needed in bare noun naming; see for a similar argumentation regarding grammatical gender La Heij et al., 1998). On the contrary, in our experiment participants were asked to produce noun phrases, more specifically, quantifier-classifier-noun phrase sequences. As a result, it was essential for participants to explicitly produce the classifiers given the target lemma. Due to this difference in the nature of

tasks, the discrepancy in results from the two studies does not necessarily point toward conflicting conclusions. It is possible that just like the case with Dutch and German and their grammatical gender effects, Mandarin Chinese is a language where classifier congruency effects can be found only in NP production tasks but not in bare noun naming. If this is true, together with conclusions of the Wang et al. (2019) study, our findings follow what was found for Germanic languages (La Heij et al., 1998; Schiller & Caramazza, 2003; Schriefers, 1993). This NP production process can be interpreted as an amplifier for the potential classifier congruency effect that is not detected in easier tasks where no explicit use of classifier is necessary.

Secondly, we would like to draw the readers' attention to the choice of stimuli in the two studies. Due to the bare noun naming task of the Wang et al. (2019) study, it was impossible to confirm with the participants whether the classifiers activated by participants were the ones of which the usage was assumed by the experimenters. Also, we were reluctant to include some of the stimuli from the Wang et al. study due to

reasons such as visual complexity, multiple available classifiers and repetition. Lastly, the difference in statistical method may also have caused minor variations in results (see [Appendix B](#)). Given these differences between the two studies, we should be careful when comparing and interpreting the results of these studies, bearing in mind the above-mentioned factors and variables that can potentially weaken our reasoning and conclusions.

Besides the significant effect of classifier congruency on naming latency discovered in the behavioral data, we also noticed in examination of fronto-central and centro-parietal regions an N400-like effect in time window 385–585 ms that was significantly larger in the classifier-incongruent than in the classifier-congruent condition. This result is consistent with the [Wang et al. \(2019\)](#) study which, albeit obtaining no difference in naming latency between classifier-congruent and incongruent conditions, still found an ERP wave that was significantly more negative in the latter than in the former condition. The observation of a stronger or weaker N400 effect by the manipulation of classifiers is a result also found in existing studies on classifier-noun inconsistency (e.g., [Tsai et al., 2008](#); [Zhou et al., 2010](#); [Zhang et al., 2012](#); [Hsu et al., 2014](#)), as well as in research studying the determiner and gender disagreement effects in Indo-European languages (e.g., [Barber & Carreiras, 2005](#)).

As demonstrated above, there are a number of similarities between classifiers and grammatical gender. Both are used to classify the nominal system, and [Dixon \(1986\)](#) has listed criteria opposing gender and classifiers. However, more recently, [Fedden and Corbett \(2017\)](#) stated that the gender-classifier division cannot be maintained because – contrary to a long-held belief in typology – some languages have both gender and classifiers, i.e. they have concurrent systems to classify their nominal system, and, moreover, classifiers can grammaticalize into gender systems. We claimed that our current study complements and extends on grammatical processing in gender-marking languages. Both gender and classifiers are lexico-syntactic features, presumably activated and selected through a lexical entry, and we have revealed some parallels regarding their processing, e.g. as mentioned above, no congruency effect when the feature is not needed. Models of language production should include classifiers as lexico-syntactic features to account for processing of a broader range of languages.

The remaining question is at which stage the classifier feature receives activation and its selection takes place. The first thing to clarify is that since the mapping of the classifier feature is quite opaque, it is the individual nouns themselves rather than their semantic categories that are the dominant and decisive factor for classifier choice ([Tzeng et al., 1991](#); [Shao, 1993](#)). The correspondence between nouns and classifiers must be memorized, and native Mandarin Chinese speakers acquire the syntactic structure and basic combinations at around three to five years of age (e.g., [Erbaugh, 1986, 2002](#); [Fang, 1985](#)). In other words, there is no good reason to presume that classifiers receive activation from semantic encoding. The possibility of activation from phonological representation is also unlikely. Therefore, the joint evidence from the two studies is compatible with [Levelt et al.'s model \(1999\)](#). According to this model, the lexico-syntactic nature of features such as grammatical gender and classifier determines that they receive activation from the encoding of the lemma stratum, which soundly explains why the classifier congruency effect in the behavioral data analysis is absent in bare noun naming but present in NP production. In order to perform a task as simple as bare noun naming, by the time activation reaches the classifier

following the retrieval of the lemma, the preparation for utterance production is already moving on towards the word form or phonological encoding stage. Since classifier information is not relevant for the production of bare picture names, there is no need to further process it, and consequently the naming speed is not affected by whether or not the classifiers of target and distractor are congruent. However, when it comes to NP production, the classifier information is necessary for carrying out the task and its phonological form must be encoded for production of the entire utterance, which is why the selection speed difference caused by the classifier congruency effect has an impact on the naming latencies of target pictures.

6. Conclusion

The current paper explored the activation/selection of Chinese classifiers in noun phrase production through an overt picture naming experiment using the picture-word interference paradigm. The behavioral analysis of naming latency shows that participants took significantly less time to respond in semantically unrelated and classifier-congruent conditions than in semantically related and classifier-incongruent conditions. Analysis of ERP data showed that the N400 component is significantly more negative in the semantically unrelated than in the related condition, and in the classifier-incongruent relative to the congruent condition. No interaction of the two factors was found in either the behavioral or ERP data analysis. Given these results, we propose that classifier as a lexico-syntactic feature goes through competitive selection in NP production. This finding complements the observations from previous work by [Wang et al. \(2019\)](#), who reported automatic activation of classifiers in bare noun naming based on the absence of a difference in naming latency relative to classifier congruency status but a stronger N400 component in the incongruent relative to the congruent condition. In terms of accounting for the specific process of classifier activation in both scenarios, [Levelt et al.'s model \(1999\)](#) provides a straight-forward account based on the serial procedure of speech production.

Our data complement and extend the previous results obtained in bare noun naming to cover a broader scope of Mandarin Chinese speech production, which in turn examines the generalizability of earlier hypotheses made on determiner and grammatical gender effect and enables further cross-language comparison by including a Sino-Tibetan language into the established discussion of Indo-European languages. Future research may want to consider examining other languages with classifier as a lexico-syntactic feature to verify whether the conclusions for Mandarin Chinese hold within a wider scope.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Jos Pacilly for his contributions to the Praat script and Leticia Pablos Robles for her advice on the statistical analysis. We also thank LUCL for partially sponsoring the remuneration of the participants.

Appendix A. Table of stimuli used in the experiment

Target picture	Classifier	Distractor Type			
		Semantically Related		Semantically Unrelated	
		C+	C-	C+	C-
兔子 tu4zi0 rabbit	只 zhi1	老鼠 lao3shu3 mouse	水牛 shui3niu2 buffalo	手掌 shou3zhang3 palm	肉 rou4 meat
刀 dao1 knife	把 ba3	叉子 cha1zi0 fork	盘子 pan2zi0 plate	扇子 shan4zi0 fan	雪茄 xue3jia1 cigar
围巾 wei2jin1 scarf	条 tiao2	裤子 ku4zi0 trousers	西装 xi1zhuang1 suit	金鱼 jin1yu2 goldfish	脸盆 lian3pen2 washbasin
树叶 shu4ye4 treeleaf	片 pian4	花瓣 hua1ban4 petal	枝丫 zhi1ya1 branch	沼泽 zhao3ze2 swamp	尸体 shi1ti3 dead body
小提琴 xiao3ti2qin2 violin	把 ba3	吉他 ji2ta0 guitar	钢琴 gang1qin2 piano	锁 suo3 lock	烟袋 yan1dai4 tobacco pipe
蚊子 wen2zi0 mosquito	只 zhi1	蝴蝶 hu2die2 butterfly	蜈蚣 wu2gong1 centipede	喇叭 la3ba1 trumpet	手枪 shou3qiang1 gun
砖头 zhuan1tou0 brick	块 kuai4	钢板 gang1ban3 steel plate	石子 shi2zi3 pebble	肥皂 fei2zao4 soap	瓜子 gua1zi3 melon seed
饺子 jiao3zi0 dumpling	个 ge4	包子 bao1zi0 steamed bun	油条 you2tiao2 dough stick	箱子 xiang1zi0 box	宝石 bao3shi2 gem
床 chuang2 bed	张 zhang1	席子 xi2zi0 mat	枕头 zhen3tou0 pillow	卡片 ka3pian4 card	雨水 yu3shui3 raindrop
蛋糕 dan4gao1 cake	块 kuai4	饼干 bing3gan1 biscuit	糖果 tang2guo3 candy	手表 shou3biao3 watch	旅馆 lv3guan3 hotel
钻石 zuan4shi2 diamond	颗 ke1	珍珠 zhen1zhu1 pearl	玉 yu4 jade	星星 xing1xing0 star	旗子 qi2zi0 flag
教堂 jiao4tang2 church	座 zuo4	宫殿 gong1dian4 palace	洋房 yang2fang2 western house	塑像 su4xiang4 statue	相机 xiang4ji1 camera
花 hua1 flower	朵 duo3	蘑菇 mo2gu0 mushroom	稻草 dao4cao3 straw	云 yun2 cloud	城堡 cheng2bao3 castle
手指 shou3zhi3 finger	根 gen1	脚趾 jiao3zhi3 toe	膝盖 xi1gai4 knee	电线 dian4xian4 wire	罐头 guan4tou can
戒指 jie4zhi0 ring	枚 mei2	胸针 xiong1zhen1 brooch	金冠 jin1guan4 crown	勋章 xun1zhang1 medal	火炬 huo3ju4 torch
蛇 she2 snake	条 tiao2	龙 long2 dragon	猪 zhu1 pig	毛巾 mao2jin1 towel	桥梁 qiao2liang2 bridge
山 shan1 mountain	座 zuo4	悬崖 xuan2ya2 cliff	河流 he2liu2 river	村庄 cun1zhuang1 village	炉子 lu2zi0 furnace
沙滩 sha1tan1 beach	片 pian4	海洋 hai3yang2 sea	岛 dao3 island	雪花 xue3hua1 snowflake	医院 yi1yuan4 hospital
耳机 er3ji1 headphones	副 fu4	眼镜 yan3jing4 glasses	头巾 tou2jin1 headcloth	对联 dui4lian2 scroll couplet	麻雀 ma2que4 sparrow
纸 zhi3 paper	张 zhang1	照片 zhao4pian4 photograph	画 hua4 painting	桌子 zhuo1zi0 table	陨石 yun3shi2 meteorite
闪电 shan3dian4 thunder	道 dao4	彩虹 cai3hong2 rainbow	乌云 wu1yun2 dark cloud	疤痕 ba1hen2 scar	峡谷 xia2gu3 valley
布 bu4 cloth	匹 pi3	绸缎 chou2duan4 silk	丝线 si1xian4 thread	骏马 jun4ma3 steed	绵羊 mian2yang2 sheep
硬币 ying4bi4 coin	枚 mei2	铜钱 tong2qian2 copper cash	钞票 chao1piao4 bill	徽章 hui1zhang1 badge	甲板 jia3ban3 deck
大米 da4mi3 rice	粒 li4	芝麻 zhi1ma0 sesame	红薯 hong2shu3 sweet potato	沙子 sha1zi0 sand	凤凰 feng4huang4 phoenix
柿子 shi4zi0 persimmon	个 ge4	橘子 ju2zi0 orange	杨梅 yang2mei bayberry	燕子 yan2zi0 swallow	苍蝇 cang1ying fly

(continued on next page)

(continued)

Target picture	Classifier	Distractor Type			
		Semantically Related		Semantically Unrelated	
		C+	C-	C+	C-
shi4zi0 persimmon 隧道	ge4 条	ju2zi0 orange 马路	yang2mei2 bayberry 车站	yan4zi0 swallow 蚯蚓	cang1ying0 fly 笼子
sui4dao4 tunnel 黄瓜	tiao2 根	ma3lu4 avenue 葱	che1zhan4 station 白菜	qiuliyin3 earthworm 拐杖	long2zi0 cage 饭碗
huang2gua1 cucumber 衬衫	gen1 件	cong1 shallot 外套	bai2cai4 Chinese cabbage 披肩	guai3zhang4 stick 行李	fan4wan3 bowl 帐篷
chen4shan1 t-shirt 井	jian4 口	wai4tao4 overcoat 缸	pi1jian1 tippet 烟囱	xing2li3 luggage 棺材	zhang4peng2 tent 字典
jing3 well 斧头	kou3 把	gang1 vat 剑	yan1cong0 chimney 箭	guan1cai2 coffin 钥匙	zi4dian3 dictionary 喜鹊
fu3tou2 axe	ba3	jian4 sword	jian4 arrow	yao4shi0 key	xi3que4 magpie

Appendix B. Analysis of variance for behavioral data on naming latency. Analysis was run using R (R Core Team, 2012) using the *ez* package (Lawrence & Lawrence, 2016). Tables are presented in the order of by-subject and by-item analysis.

Source	F		
	DFn	DFd	
Classifier Congruency	1	20	4.48*
Semantic Relatedness	1	20	10.48*
Classifier × Semantic	1	20	0.23

*p < 0.05.

**p < 0.01.

***p < 0.001.

Source	F		
	DFn	DFd	
Classifier Congruency	1	29	4.11
Semantic Relatedness	1	29	3.9
Classifier × Semantic	1	29	1.07

*p < 0.05.

**p < 0.01.

***p < 0.001.

In the by-subject ANOVA, both classifier congruency and semantic relatedness had a significant effect on naming latency ($p = 0.047$; $p = 0.004$). No significant difference was observed from the interaction between the two factors. In the by-item ANOVA, however, neither classifier congruency nor semantic relatedness passed the significance threshold, although p-values were very close to an alpha level of 0.05 ($p = 0.052$; $p = 0.058$). These results, including the marginal p-values in the by-item analysis, indicate that main effects of both classifier congruency and semantic relatedness were significant on naming latency. In other words, naming latency was significantly shorter in the classifier-congruent than in the incongruent condition, as well as in the semantically unrelated than in the related condition.

The outcome of our experiment is different from that of the Wang et al. (2019) study that did not find classifier congruency effects in either analysis but significant semantic relatedness effect in both, which is explained by the different task requirements and hence different production processes in the two studies (see Section 5). However, the absence of a significant effect in the by-item ANOVA in our study points to a potential defect that can be associated with running ANOVA with condition means obtained for each item. To justify such an approach and the fact that significant results were obtained only in the by-subject analysis, we must assume that items used in the experiment can be treated as fixed factors (Clark, 1973), while in our case (and in the Wang et al. study as well) there is essentially no sound ground for this assumption. Given that both our participants and items were randomly drawn from a vast population, we would prefer to control for the randomness and avoid the risk of increasing a Type I error from item/subject variability in F1 and F2 tests. As a result, we adopted the general mixed effects model in our study for data analysis instead of an ANOVA. Nonetheless, ANOVA results are still included for the purpose of comparison with the behavioral analysis in Wang et al. (2019) study.

References

- Alario, F. X., & Caramazza, A. (2002). The production of determiners: Evidence from French. *Cognition*, 82, 179–223.
- Allan, K. (1977). Classifiers. *Language*, 53, 285–311.
- Barber, H., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience*, 17, 137–153.
- Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using Eigen and R syntax. R package version 0.999999-0.
- Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology*, 58, 667–692.
- Bi, Y., Yu, X., Geng, J., & Alario, F. X. (2010). The role of visual form in lexical access: Evidence from Chinese classifier production. *Cognition*, 116, 101–109.
- Bloem, I., & La Heij, W. (2003). Semantic facilitation and semantic interference in word translation: Implications for models of lexical access in language production. *Journal of Memory and Language*, 48, 468–488.
- Boersma, P., & Weenink, D. Praat: Doing Phonetics by Computer. 2019. [Computer program].

- Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14, 177–208.
- Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the tip-of-the-tongue phenomenon. *Cognition*, 64(3), 309–343.
- Caramazza, A., Miozzo, M., Costa, A., Schiller, N. O., & Alario, F.-X. (2001). A cross-linguistic investigation of determiner production. In E. Dupoux (Ed.), *Language, brain and cognitive development: Essays in honor of Jacques Mehler* (pp. 209–226). Cambridge, MA: MIT Press.
- Chen, J.-Y., & Wang, T.-Y. (2003). The nature of the classifier–noun agreement in Chinese word production. Paper presented at the 44th annual meeting of the Psychonomic Society, Vancouver, Canada.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- Costa, A., Sebastián-Gallés, N., Miozzo, M., & Caramazza, A. (1999). The gender congruity effect: Evidence from Spanish and Catalan. *Language and Cognitive Processes*, 14, 381–391.
- Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, 106(50), 21442–21446.
- Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45, 145–171.
- Cubelli, R., Lotto, L., Paolieri, D., Girelli, M., & Job, R. (2005). Grammatical gender is selected in bare noun production: Evidence from the picture–word interference paradigm. *Journal of Memory and Language*, 53, 42–59.
- Dell'Acqua, R., Sessa, P., Peressotti, F., Mulatti, C., Navarrete, E., & Grainger, J. (2010). ERP evidence for ultra-fast semantic processing in the picture–word interference paradigm. *Frontiers in Psychology*, 1, 177.
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 283–321.
- Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, 27, 124–142.
- Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, 5, 313–349.
- Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review*, 98, 604–614.
- Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42, 287–314.
- Dhooge, E., De Baene, W., & Hartsuiker, R. J. (2016). The mechanisms of determiner selection and its relation to lexical selection: An ERP study. *Journal of Memory and Language*, 88, 28–38.
- Dixon, R. M. W. (1986). Noun classes and noun classification in typological perspective. In Craig, C. (Ed.), *Noun classes and categorization: Proceedings of a symposium on categorization and noun classification*, Eugene, Oregon, October 1983 (pp. 105–112). Amsterdam: John Benjamins.
- Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In Craig, C. (Ed.), *Noun classes and categorization: Proceedings of a symposium on categorization and noun classification*, Eugene, Oregon, October 1983 (pp. 399–436). Amsterdam: John Benjamins.
- Erbaugh, M. S. (2002). Classifiers are for specification: Complementary functions for sortal and general classifiers in Cantonese and Mandarin. *Cahiers de linguistique Asie orientale*, 31, 33–69.
- Fang, F. (1985). An experiment on the use of classifiers by 4- to 6-year-olds. *Acta Psychologica Sinica*, 17, 384–392.
- Fedden, S., & Corbett, G. G. (2017). Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: A Journal of General Linguistics*, 2(34), 1–47.
- Finocchiaro, C., Alario, F.-X., Schiller, N. O., Costa, A., Miozzo, M., & Caramazza, A. (2011). Gender congruency goes Europe: A cross-linguistic study of the gender congruency effect in Romance and Germanic languages. *Italian Journal of Linguistics*, 23, 161–198.
- Friederici, A. D., & Jacobsen, T. (1999). Processing grammatical gender during language comprehension. *Journal of Psycholinguistic Research*, 28, 467–484.
- Ganushchak, L., Christoffels, I., & Schiller, N. O. (2011). The use of electroencephalography in language production research: A review. *Frontiers in Psychology*, 2, 208.
- Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions* (pp. 19–76). London: Academic Press.
- Glaser, W. R. (1992). Picture naming. *Cognition*, 42, 61–105.
- Glaser, W. R., & Döngelhoff, F. J. (1984). The time course of picture–word interference. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 640–654.
- Heim, S., Friederici, A. D., Schiller, N. O., Rüschemeyer, S.-A., & Amunts, K. (2009). The determiner congruency effect in language production investigated with functional MRI. *Human Brain Mapping*, 30, 928–940.
- 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.
- Janssen, N., Schiller, N. O., & Alario, F.-X. (2014). The selection of closed-class elements during language production: A reassessment of the evidence and a new look on new data. *Language, Cognition and Neuroscience*, 29, 695–708.
- Koester, D., & Schiller, N. O. (2008). Morphological priming in overt language production: Electrophysiological evidence from Dutch. *Neuroimage*, 42, 1622–1630.
- 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.
- La Heij, W. (1988). Components of Stroop-like interference in picture naming. *Memory & Cognition*, 16, 400–410.
- La Heij, W., Mak, P., Sander, J., & Willeboordse, E. (1998). The gender-congruency effect in picture-word tasks. *Psychological Research*, 61, 209–219.
- Lakoff, G. (1986). Classifiers as a reflection of mind. *Noun Classes and Categorization*, 7, 13–51.
- Lapointe, S. G., & Dell, G. S. (1989). A synthesis of some recent work in sentence production. In G. N. Carlson, & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 107–156). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez'. R package version, 4–4.
- Leckey, M., & Federmeier, K. D. (2019). Electrophysiological Methods in the Study of Language Processing. In G. I. de Zubicaray & N. O. Schiller (Eds.), *The Oxford Handbook of Neurolinguistics* (pp. 42–71). Oxford, New York: Oxford University Press.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–38.
- Levelt, W. J. M., & Schriefers, H. (1987). Stages of lexical access. In G. Kempen (Ed.), *Natural language generation* (pp. 395–404). Springer, Dordrecht: Martinus Nijhoff Publishers.
- Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). Normal and deviant lexical processing: Reply to Dell and O'Seaghdha (1991). *Psychological Review*, 98, 615–618.
- Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. University of California Press.
- Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, 54, 146–157.
- Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture–word interference. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 36, 349–367.
- Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture–word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 503–535.
- Miozzo, M., & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 907–922.
- Nickels, L., Biedermann, B., Fieder, N., & Schiller, N. O. (2015). The lexical-syntactic representation of number. *Language, Cognition and Neuroscience*, 30, 287–304.
- Qian, G., Reinking, D., & Yang, R. (1994). The effects of character complexity on recognizing Chinese characters. *Contemporary Educational Psychology*, 19, 155–166.
- R Core Team (2012). R: A language and environment for statistical computing. 2012. Vienna, Austria: R Foundation for Statistical Computing, 10.
- Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42, 107–142.
- Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, 47, 59–87.
- Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, 64, 249–284.
- Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, 110, 88–125.
- Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, 18, 367–379.
- Scalise, S. (1994). *Le strutture del linguaggio. Morfologia*. Bologna: Il Mulino.
- Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, 48, 169–194.
- Schiller, N. O., & Caramazza, A. (2006). Grammatical gender selection in language production: The case of diminutives in Dutch. *Language and Cognitive Processes*, 21, 945–973.
- Schiller, N. O., & Costa, A. (2006). Different selection principles of free-standing and bound morphemes in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1201–1207.
- Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 841–850.
- Shao, J. (1993). 量词的语义分析及其与名词的双向选择. *中国语文*, 3(289), 2.
- Shi, Y. Z. (1996). Proportion of extensional dimensions: The primary cognitive basis for shape-based classifiers in Chinese. *Journal-Chinese Language Teachers Association*, 31, 37–60.
- Starreveld, P., & La Heij, W. (2004). Phonological facilitation of grammatical gender retrieval. *Language and Cognitive Processes*, 19, 677–711.
- Stemberger, J. P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology*, 1, 281–313.
- Strijkers, K., Holcomb, P. J., & Costa, A. (2011). Conscious intention to speak proactively facilitates lexical access during overt object naming. *Journal of Memory and Language*, 65, 345–362.
- Tai, J. H. (1994). Chinese classifier systems and human categorization. In M. Chen, & O. Tzeng (Eds.), *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change* (pp. 479–494). Pyramid Publishing Co.
- Tai, J. H., & Chao, F. Y. (1994). A semantic study of the classifier zhang. *Journal of the Chinese Language Teachers Association*, 29, 67–78.
- Tai, J., & Wang, L. (1990). A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association*, 25, 35–56.

- Tien, Y. M., Tzeng, O. J., & Hung, D. L. (2002). Semantic and cognitive basis of Chinese classifiers: A functional approach. *Language and Linguistics*, 3, 101–132.
- Tsai, S.-H. R., Hsu, C.-C. N., Yang, C.-L., & Chen, J.-Y. (2008). An event-related potential (ERP) study of the classifier–noun relationship in Mandarin Chinese. *Paper presented at the British Association of Cognitive Neuroscience (BACN) joint annual meeting with the Wales Institute for Cognitive Neuroscience (WICN)*.
- Tzeng, O. J., Chen, S., & Hung, D. L. (1991). The classifier problem in Chinese aphasia. *Brain and Language*, 41, 184–202.
- Van Berkum, J. J. (1997). Syntactic processes in speech production: The retrieval of grammatical gender. *Cognition*, 64, 115–152.
- Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, 38, 584–589.
- 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, 131–150.
- Wang, L., Guo, J., Bi, Y., & Shu, H. (2006). Classifier congruency effect in the production of noun phrases. *Studies of Psychology and Behavior*, 4, 34–38.
- Wang, M., Chen, Y., & Schiller, N. O. (2019). Lexico-syntactic features are activated but not selected in bare noun production: Electrophysiological evidence from overt picture naming. *Cortex*, 116, 294–307.
- Wang, M., & Schiller, N. O. (2019). A review on grammatical gender agreement in speech production. *Frontiers in Psychology: Language Sciences*, 9, 2754.
- Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex*, 39, 483–508.
- Zhang, J. J., & Liu, H. Y. (2009). The lexical access of individual classifiers in language production and comprehension. *Acta Psychologica Sinica*, 41, 580–593.
- 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, 321–331.
- 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, 1551–1562.
- Zhu, X., Damian, M. F., & Zhang, Q. (2015). Seriality of semantic and phonological processes during overt speech in Mandarin as revealed by event-related brain potentials. *Brain and Language*, 144, 16–25.