The lexico-semantic representation of words in the mental lexicon = De lexico-semantische representatie van woorden in het mentale lexicon

Wang, Y.

CHAPTER 4

Activation of classifiers in word production: insights from lexico-syntactic probability distributions

This article is published as: Wang, Y., Witteman, J., & Schiller, N. O. (2025). Activation of classifiers in word production: insights from lexico-syntactic probability distributions. Language, Cognition and Neuroscience,40(7), 987–1001.

Abstract: In speech production, speakers need to activate and select appropriate lexico-syntactic features of words to plan grammatically correct sentences. While previous studies have focused on situations with a single grammatically correct option, Mandarin Chinese classifiers present a case where multiple options are correct. In this study, we asked native Mandarin Chinese speakers to name target pictures using a picture-word interference paradigm while recording EEG. Distractor words with varying degrees of classifier distribution similarity were superimposed while controlling for dominant classifier congruency. Distractors with dissimilar classifier distributions resulted in a more positive P600-like effect, but no behavioural effect was observed, compared to distractor nouns having similar classifier distributions. Based on this result, we propose that, when producing a bare noun, multiple compatible classifiers are activated at the lexical level with the degree of activation being determined by their corresponding compatibility with the given noun.

Keywords: Language production; classifier probability distributions; Jensen Shannon divergence (JSD); Mandarin Chinese classifiers; picture-word interference (PWI); bare noun naming; P600

4.1 Introduction

To produce grammatically correct sentences, speakers must retrieve lexico-syntactic features such as grammatical gender or classifiers for a given noun. The number of grammatically correct options for a lexico-syntactic feature can be either single or multiple. For instance, grammatical gender of nouns in many Indo-European languages has a single grammatically correct option, e.g. Haus (‘house”) is neuter in German (not feminine or masculine) and maison is feminine in French (not masculine), etc. In contrast, multiple grammatically correct options are also common, for instance, in the case of classifiers for nouns in Mandarin Chinese. Specifically, Mandarin Chinese classifiers are a grammatically mandated component of a noun when expressing quantities or definitions of an entity. This function is realised via the “this/that + classifier + noun” structure.

The exact choice of the classifier in this setting is, to some extent, determined by the intended meaning and/or semantic features of the object (Lakoff, 1986; Tai, 1994; Wu & Bodomo, 2009). Classifiers perform various functions. For instance, sometimes the classifier is used as a means to count a noun, such as in the case of e.g. 三本书(/san1ben3shu1/, three + CL: “unit of a book” + books). Or the classifier serves to specify a unit of measurement such as in the case of, e.g. 三杯水(/san1bei1shui3/, three + CL: cups [of] + water). Alternatively, a classifier plays a role as a descriptor of the shape of a noun, such as in the case of 一条蛇(/yi4tiao2she2/, a + CL: longness + snake). Importantly, a given noun is (typically) compatible with multiple classifiers, though (grammatically speaking) only one classifier can be used at a time. Between the various compatible classifiers for a given noun, each classifier serves to emphasise a different perspective of the given noun. For instance, 邮票(/you2piao4/, stamp; digits refer to different tones, i.e. 1-4) is “flat and thin” and “tiny”, which makes it compatible with both 张(/zhang1/, i.e. the classifier associated with long and thin entities) and 枚(/mei2/, i.e. the classifier associated with tiny entities) (Wu & Bodomo, 2009). Hence, both 那张邮票(“that zhang1 stamp") and 那枚邮票("that mei2 stamp") are grammatically correct. However, corpus research shows that 张(/zhang1/) is the most common choice of classifier for 邮票(/you2piao4/, stamp), making it its dominant classifier. The dominant classifier, together with all other possible classifiers and their probabilities of co-occurrence, yields a classifier probability distribution for a given noun. Importantly, the probability of classifiers for a given noun can take the value of 0, denoting full incompatibility of a classifier with the given noun, i.e. grammatically incorrect classifier for the given noun (Liu et al., 2019; Wu & Bodomo, 2009). A value of 1 would indicate full compatibility of the classifier with the noun.

Studies regarding Mandarin Chinese classifiers have thus far been simplified to dichotomous situations where only the dominant classifier is presumed to be correct (e.g., Huang & Schiller, 2021; Wang et al., 2019; Wang et al., 2024). In part, this is done based on existing language production models not accounting for situations where multiple grammatically correct options for lexico-syntactic features exist. For example, in Levelt's model, language production is divided into three sequential strata: the conceptual stratum, the lemma stratum, and the phonological word-form stratum (Levelt et al., 1999; Roelofs, 1992). In order to produce a bare noun or a noun phrase (NP), speakers first conceptualise the meaning at the conceptual level, then the meaning becomes lexicalised at the lemma level before finally being articulated. At the stage of lexicalisation, the corresponding lexico-syntactic features of a given noun (e.g. gender, number, case, etc.) will be activated and subsequently selected when needed for the task. This activation and selection process is realised through a unidirectional flow of activation from the lexicalised concept to the grammatically correct option of the corresponding lexico-syntactic features (Levelt et al., 1999; Roelofs, 1992). For instance, when a speaker intends to produce the German form *Hauses*, the features neuter gender, singular number, and genitive case become activated.

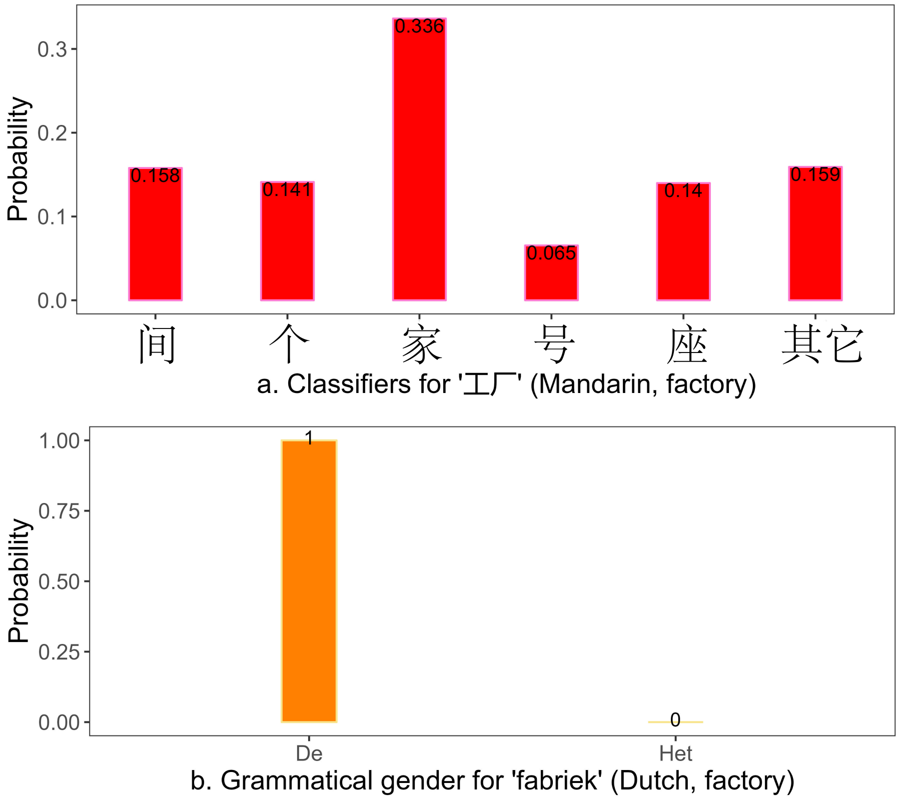
Experimentally, the activation and selection procedure of the single grammatically correct option for the lexico-syntactic feature during language production is manifested at the behavioural level. For instance, in the picture-word interference (PWI) paradigm, naming latencies are shorter when the participants are asked to name a picture with a determiner NP such as *de appel* ("the apple" in Dutch) whilst being presented with a distractor noun under congruent vs. incongruent grammatical gender conditions (Schiller & Caramazza, 2003; Schriefers, 1993). However, when the activated lexico-syntactic information is not needed for the task, such as in bare noun naming where no gender-marked items are produced, the corresponding lexico-syntactic features will not be selected. This is reflected by the absence of a detectable behavioural effect of grammatical gender congruency in the PWI paradigm (La Heij et al., 1998).

In theory, in the absence of any behavioural effects, lexico-syntactic features could still be activated at the lemma level. Such activation without selecting a lexico-syntactic feature can be observed at the electrophysiological level. For instance, Wang et al. (2019) reported that in the PWI paradigm, naming of bare nouns in Mandarin Chinese resulted in a significantly more negative N400 component for dominant classifier incongruent vs. congruent conditions, although no significant difference was observed at the behavioural level. This result is consistent with sentence comprehension studies which manipulated the classifier-noun congruency and observed an N400, suggesting that Chinese classifier-noun integration was primarily semantically driven (Qian & Garnsey, 2015). In our own previous study (i.e., Wang et al., 2024), we observed a more positive P600 effect instead for the same conditions as tested in Wang et al. (2019), similarly without any behavioural effects. In that study, we hypothesised that the difference in the observed event-related components between the two studies (the N400 effect in Wang et al., 2019 vs. the P600 effect in Wang et al., 2024 arises from a difference in the underlying classifier processing mechanism (semantically vs. lexico-syntactically driven activation) due to the varying degrees of influence between semantic features and classifier congruency in these two studies. Importantly, both Wang et al. (2024) and Wang et al. (2019) came to the same conclusion that the dominant classifier is activated but not selected. However, in the absence of co-varying semantic features, classifier congruency could elicit a lexico-syntactically driven P600.

Although the current literature has so far only explored single grammatically correct options for lexico-syntactic features, in Mandarin Chinese (as outlined above) a given noun can, based on corpus research, be thought as having a classifier *distribution* (illustrated in Figure 4.1a). Such a distribution is defined as the probability of co-occurrence between the choice of a lexico-syntactic feature and its associated noun. The probability of each option in this distribution can be seen as their extent of compatibility with the given noun. That is, for a given noun, multiple classifiers are compatible (grammatically correct) in Mandarin Chinese. However, some are used more often than others, and the most often used one is called the dominant classifier. The other non-dominant classifiers have varying degrees to which they are used with the noun. Hence, all together, for a given noun, there is a probability distribution of compatible classifiers.

This distribution can also be thought of as a technically generic version of the single grammatically correct case for lexico-syntactic features. That is, for the single grammatically correct option, the given noun always co- occurs with the grammatically correct option in an NP, and thus the probability of the corresponding option is 1 (i.e. fully compatible - see Figure 4.1b for illustration) whereas grammatically incorrect options have probabilities of 0 (i.e. fully incompatible). Thus, this approach of representing the probability of grammatically correct option(s) for a lexico-syntactic feature co-occurring with a given noun can accommodate both single and multiple grammatically correct options. Thus, we posit that such a lexico-syntactic probability distribution could more precisely account for the encoding of lexico-syntactic features in language production.

Such hypothetical encoding of multiple grammatical options for a lexico-syntactic feature in language production can be experimentally tested with the PWI paradigm by manipulation of the similarities between probability distributions of the classifier(s) of the target and distractor nouns. To achieve this, the Jensen- Shannon divergence (JSD) metric was used to quantify the degree of dissimilarity between two distributions. Values of the JSD metric range between 0 and 1, with larger values denoting more dissimi-



a. Classifiers for '工厂' (Mandarin, factory)

b. Grammatical gender for 'fabriek' (Dutch, factory)

Figure 4.1: The lexico-syntactic feature probability distributions for the a) Chinese and b) Dutch nouns translating to factory.

larity/divergence between two probability distributions. The JSD is calculated according to Equation 4.1 (Menéndez et al., 1997; Nielsen, 2020). For instance, here we have three different probability distributions of classifiers i.e. for noun A {ge: 0, zhi:0.3, tiao: 0.7}, noun B {ge: 0.7, zhi:0, tiao: 0.3}, and noun C {ge: 0.1, zhi:0.2, tiao: 0.7}. Noun A and B have the same probabilities distributed over different classifiers, whereas noun A and C only differ in the probability of classifier "zhi" and "ge". The similarity captured by JSD between noun A and B is lower than between noun A and C. Thus, in the current study, when we obtain the classifier probability distribution for any given noun, we included in all classifiers in the classifier probability distribution for the given noun by assigning fully incompatible classifiers a probability of zero.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{4.1}$$

Where P and Q are classifier probability distributions for two nouns, respectively.

$D(P||M) = \sum_i P(i)\log\frac{P(i)}{M(i)}, D(Q||M) = \sum_i Q(i)\log\frac{Q(i)}{M(i)}, M = \frac{1}{2}(P+Q)$

4.1.1 The current study

In the current study, we investigated the Mandarin Chinese classifier as an example of a language where there are multiple grammatically correct options in bare noun naming, which is in contrast to the existing literature about the lexico-syntactic features with respect to gender where typically only one grammatically correct option exists. This allows us to investigate to what extent models that have been formulated based on gender agreement also can account for more complex speech production in Mandarin Chinese. Note that for a given noun, its classifier probability distribution also defines the dominant classifier and thus necessitates controlling for dominant classifier effects in the experimental design.

Based on Levelt's model, one would hypothesise that lexico-syntactic features are encoded by a probability distribution in language production at the lemma-level and on the electrophysiological level therefore a P600-like ERP wave would be expected for dissimilar classifier distributions as compared to similar classifier distributions, akin to the P600 found for classical syntactic violation (such as agreement). Note that this hypothesis does not contradict the standard Levelt model but is a generalisation of Levelt's model (Levelt et al., 1999; Roelofs, 1992) to cases where multiple grammatically correct options for a lexico-syntactic feature exist.

Regarding the classifier probability distribution, we expect a more positive ERP amplitude between 275–575 ms post stimulus onset for the similar vs. dissimilar classifier probability distribution (i.e. lexico-syntactically driven P600 effect), but no behavioural effects. Regarding dominant classifier effects, we manipulated dominant classifier congruency in such a way that its effect was not confounded with any possible lexico-syntactic probability distribution effects. However, we predict that the dominant classifier results in no behavioural effect, in line with previous studies (Wang et al., 2019; Wang et al., 2024) and semantically-driven processing of classifiers, i.e. a significantly more negative amplitude between 275–575 ms (i.e. a N400 effect) for dominant classifier incongruent vs. congruent conditions (Wang et al., 2019; Wang et al., 2024). The predictions were made based on the assumption that classifier compatibility operates at the lemma-level because the Mandarin Chinese classifier is a lexico-syntactic feature, which is activated at the lemma-level in Levelt's model.

4.2 Methods

4.2.1 Participants

Thirty-six native Mandarin Chinese speakers (two of which were removed from further analysis later) meeting the pre-determined eligibility criteria gave informed consent to participate in this study (Detailed information regarding the sample size justifications can be found in section 4.2.2). The inclusion criteria were as follows: aged from 18–35, having normal or corrected-to-normal vision, earned (or studying for) a university degree, and no self-reported history of neurological/psychological impairments or language disorders. Each participant received €15 as compensation for their participation in this study. Given these selection criteria, we assume that participants have sufficient knowledge to grasp the relationships between the target and distractor words at the classifier level although we did not explicitly ask them about this information. The study was approved by the ethics committee of the Faculty of Humanities at Leiden University.

4.2.2 Materials

We reused nouns from Bürki et al. (2020) by translating them into Mandarin Chinese, yielding 168 nouns in Mandarin Chinese. We used the nouns from Bürki et al. (2020) because this study can be seen as a landmark study on the effect of semantic category relationships in the literature and by using their system of categorising nouns, we transparently connect our study to the existing literature. The translations were performed using Google Translate and verified through two methods: (1) translating Chinese back to English to ensure the original word could be retrieved, and (2) manual verification by the first author. No further norming was deemed necessary with respect to semantic categories because we used the categorisation published in Bürki et al. (2020). The frequency of distractors and the distributions of classifiers for all nouns employed in the current study were determined by calculating frequency of (co-)occurrence in the Chinese Wikipedia data released on 2021- 11-01 (Wikipedia, 2024) with only taking 1-gram and 2-gram into account after word-segmentation with the library *pkuseg* in Python (Luo et al., 2019). Detailed information about the classifier probability distributions for each word in this study, the skewness of its distribution, and distractor word frequency can be found at https://github.com/Yufanggg/LexicoProbDistri.

Based on these distributions, we first calculated the JSD value for each word pair. Then, we only selected word pairs which have extremely dissimilar (D-, i.e. JSD values greater than or equal to 0.6) or similar (D+, i.e. less than or equal to 0.4) distributions to maximise the chance of detecting the effects of probability distribution similarities (Mack, 2016). These thresholds were chosen to make sure that the D- and D+ conditions were sufficiently different in JSD value to allow for the detection of the effect with our design. The dominant classifier was defined as the classifier having the highest probability for a given noun and further validated by using the Xinhua dictionary (11th edition, Linguistics Institute of Chinese Academy of Social Sciences, 2011).

We also used the Xinhua dictionary Xinhua dictionary (11th edition, Linguistics Institute of Chinese Academy of Social Sciences,

2011) to validate the number of strokes of distractor nouns. The distractor nouns were selected based on the similarity (similar, D+, vs. dissimilar, D-) of classifier probability distributions and congruency of dominant classifier (congruent, C+, vs. incongruent, C-) with the target nouns (see Figure 4.2 for an example). The frequency of distractors was determined by calculating frequency of occurrence in the Chinese Wikipedia data released on 2021-11-01 (Wikipedia, 2024). The distractor words have similar word frequency, number of strokes, and lengths in characters across all conditions. Specifically, frequency (F = 1.202, df = (2, 99), p = 0.305), number of strokes (F = 0.137, df = (2, 99), p = 0.872), and length of distractor nouns in characters (F = 0.905, df = (2, 99), p = 0.408) showed no significant difference between conditions. Distractors in the current study had no orthographic or phonological relationship with target picture names.

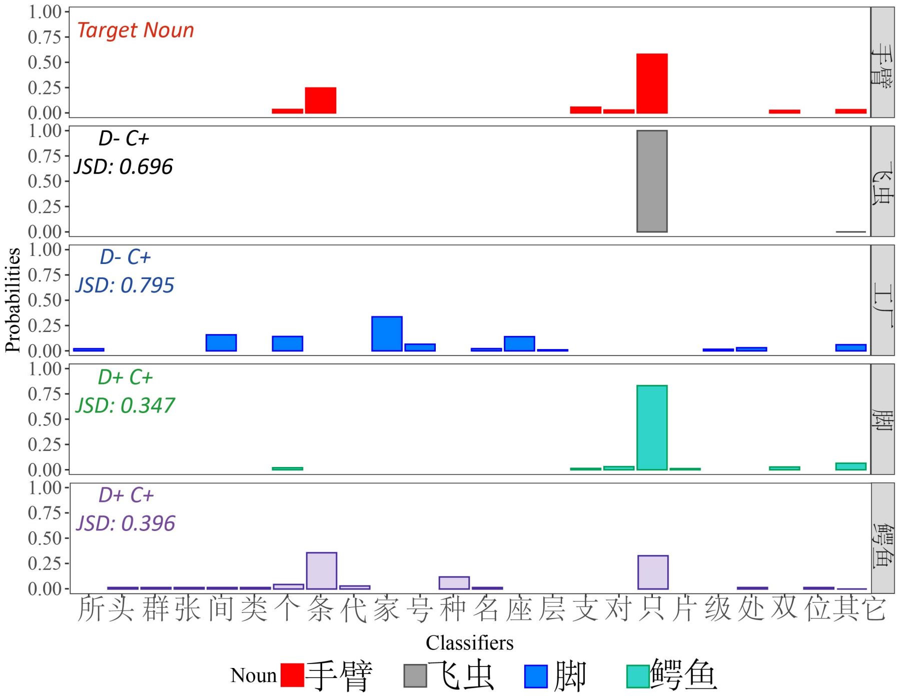


Figure 4.2: An example of target noun (手臂, /shou3bi4/, "arm") and its paired distractors for each condition ("bug" for D-C+, "factory" for D-C-, "foot" for D+C+, "alligator" for D+C-).

In the end, we selected 38 target nouns corresponding to black-and-white drawings, which we produced ourselves (see at https://github.com/Yufanggg/LexicoProbDistri). Regarding the employed target words in this study, a small number of them were chosen to be more formal (written) rather than more common (spoken) words, which might have influenced the results. However, this applied only to the following two words, i.e. 手臂(/shou3bi4/, arm), and 披风(/pi1feng1/, cloak), thus only having minimal impact. The targets and the corresponding distractor nouns used in the study, together with their properties (i.e. word frequency, number of strokes, length in characters, and the semantic congruency with the targets) also can be found at https://github.com/Yufanggg/LexicoProbDistri.

In summary, in this experiment, we manipulated the classifier probability distribution similarity (similar, D+, vs. dissimilar, D-) and dominant classifier congruency (congruent, C+, vs. incongruent, C-) with the target nouns. We also accounted for other variables, e.g. number of strokes, word frequency, and length of the distractors (Bürki et al., 2020) that might affect the outcomes of this study according to the literature. We also included the semantic category congruency effect to (1) control for semantic relatedness and (2) verify that our study design can replicate this established effect. As outlined in the introduction, no behavioural classifier effect is expected, but a semantic congruency effect is. We decided to sufficiently power our study to detect this effect, in order to (1) replicate the result (2) and enable us to detect ERP effects of similar size for classifier congruency. Therefore, the sample size was determined to detect the semantic congruency effect as found in Bürki et al. (2020), namely the semantic (category) interference effect. We performed power simulations to calculate the number of participants necessary to detect the effect. A detailed results of the simulations is provided in the Appendix 4.

4.2.3 Experimental procedure

This experiment consisted of familiarisation, practice, and experimental sessions. In the familiarisation session, the target pictures (see Figure 4.3 below) were presented with their names underneath for three seconds. Participants were required to have a thorough look at the picture and their name underneath. In the practice session, the same pictures with the string "XXX" superimposed were shown for three seconds. Participants were required to name the target pictures while ignoring the string "XXX." In the experimental session (as shown in Figure 4.3), each trial included a fixation cross ("+", 300 ms), followed by a blank screen (200 ms), a picture-distractor display (3,000 ms), and another blank screen (500 ms). Picture-distractor displays included a picture paired with a distractor word (from one of the four conditions) at the centre. Participants were required to produce the picture's name as quickly and accurately as possible while ignoring the superimposed distractor. The order of trials was pseudo-random to prevent consecutive repetitions of identical target nouns and order effects. The whole experimental session was divided into two blocks with a short break in between.

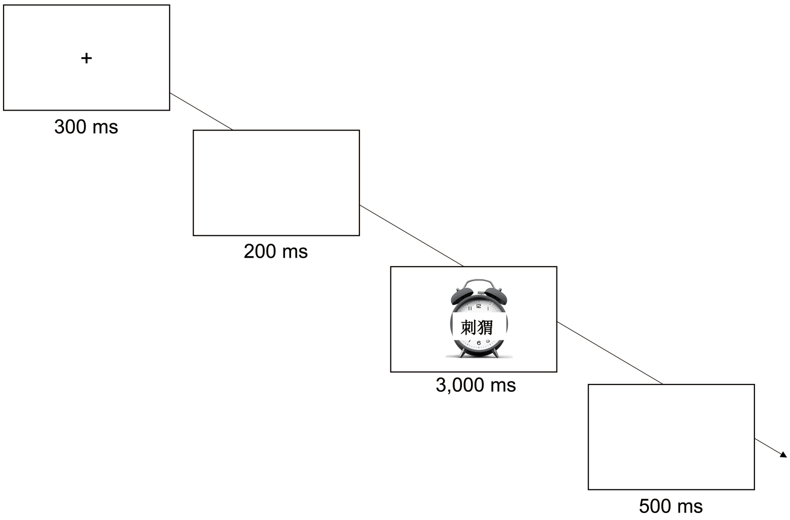


Figure 4.3: Sequence of events of each trial in the experimental session.

E-prime version 3 was used for stimulus presentation. During the experiment, participants sat in front of a computer screen in a dimly lit room and completed the task while vocal responses and the electroencephalogram (EEG) were being recorded simultaneously.

4.2.4 Audio and electroencephalography recordings

During the experiment, the vocal responses were recorded via E-prime version 3 using inline scripts and BrainVision Recorder software (version 1.23.0001) from Brain Products GmbH. Together with recording the vocal responses, the EEG was recorded using 32-channel EasyCap electrodes according to the international 10/20 system plus additional electrodes. The additional electrodes were placed above and below the left eye (VEOG) to monitor vertical eye movements and on the external canthus of each eye (HEOG) for the horizontal electrooculogram and on the mastoids. Impedance during the experiment was controlled and kept below 5 kΩ, and the sampling rate was 1,024 Hz using actiCAP control software (version 1.2.5.3).

4.2.5 Data analysis

Behavioural data analysis

Trials with incorrect vocal responses or latencies longer than 3,000 ms were regarded as incorrect trials. Naming latencies for correct trials were extracted from the sound recordings using Praat version 6.1.09 (Boersma, 2007). Trials with naming latencies larger than 3 SDs away from the individual subject and item mean were excluded (2.95% of all data, i.e. 1.17% for D-C+, 2.96% for D-C-, 3.22% for D+C+, and 4.49% for D+C-). Naming accuracies and naming latencies were analyzed with the *glmer()* function in the lmer library (version 1.1-29) in R (version 4.1.1) with the binomial and Gamma (identity) as link functions, respectively (Lo & Andrews, 2015). More specifically, items and participants were included as random factors. The word frequency of distractor words, the number of strokes in distractor nouns and length of distractor nouns in characters were first centred and then included as (fixed) nuisance factors together with the congruency of semantic relatedness (related, S+ vs. unrelated, S-) to adjust potential confounds. The congruency of the dominant classifier (congruent vs. incongruent) (sum coded, 1 vs. -1) and similarity of the classifier distribution (similar vs. dissimilar) (sum coded, -1 vs. 1) were included as fixed predictors. The maximal random effect structure was determined using a backward elimination strategy, where the BIC, AIC (Kuha, 2004), and/or approximate likelihood ratio tests (Lewis et al., 2011) were employed as criteria for model selection (Bates et al., 2014; Bates, 2007). When non-convergence and/or singular fits occurred, the random effect structure was simplified until the model does not have these issues (Barr et al., 2013). The model assumptions were checked by visualising residuals and model predicted values.

EEG data analysis

EEG data pre-processing
The MATLAB 2017b toolbox EEGLab 14_0_0b (Delorme & Makeig, 2004) was used for the off-line pre-processing of the EEG data. The pre-processing included re-referencing, band-pass filtering, notch filtering, resampling, extracting epochs, baseline correction, bad channel interpolation, visual trial rejections, removing artifacts, and trial rejection. Re-referencing was performed based on the average of both mastoid electrodes. The band-pass filter was performed from 0.1–30 Hz, and a notch filter was applied from 48 to 52 Hz to decrease power line noise interference (Ahmad et al., 2012). This ensures that the ERP effect is preserved (Zhang et al., 2024) and that the current study remains comparable with previous research (Wang et al., 2019; Wang et al., 2024). Resampling was done from 1,024–256 Hz to be comparable with previous studies (Huang & Schiller, 2021; Wang et al., 2019; Wang et al., 2024). The baseline correction was performed using the -200-0 ms pre-stimulus interval. For noisy channels, interpolation was carried out. Noisy trials were rejected based on visual inspection. Artifact rejection was performed using independent component analysis (ICA) with ADJUST v1.1.1. (Mognon et al., 2011). Finally, automatic rejection was carried out on trials with an amplitude of more than ±100 μV. Participants with more than 2/3 of the trials rejected were not included for further analysis. As a result, thirty-four of the original thirty-six participants remained for further analysis in this study.

A priori amplitude analyses in time windows
To be able to compare the results with the existing literature, we first conducted a priori analyses at the identical time windows and electrodes used in previous studies (Wang et al., 2019; Wang et al., 2024). Specifically, we included the amplitude of F3, FC1, FC5, C3, CP1, CP5, P3, PO3, F4, FC2, FC6, C4, CP2, CP6, P4, and PO4 in the 275–575 ms time window as the dependent variable. Electrodes were grouped into centro-parietal (CP1, CP5, P3, PO3, CP2, CP6,

P4, and PO4) and frontocentral regions (F3, FC1, FC5, C3, F4, FC2, FC6, and C4). Time was also mean-centred and standardised within the 275–575 ms time window and included in fixed effects in the linear mixed model. Otherwise, all modelling steps were the same as for the behavioural data analysis (see in section 4.2.5).

Exploratory permutation-based cluster mass analyses (200–700 ms)

Next, in order to capture the full temporospatial extent of the manipulated variables on the EEG, a permutation-based mass univariate cluster test was performed. First, a permutation linear mixed model with threshold-free cluster enchantment (TFCE) as Type-I error correction was conducted (E = 0.66, H = 2; see Smith & Nichols, 2009) to identify time windows and channels where an effect was present across the combined four levels of the two main effects (Visalli et al., 2024) (Visalli et al., 2024). The family-wise error for the cluster permutation test was set at 5%. Amplitude ∼ Number of strokes + Frequency of distractor + Conditions (the combined four levels of the two main effects) + (1 | participant) + (1 | item) was the formula used to conduct the permutation test to meet the criteria of exchangeability under the null hypothesis of the permutation test across the combined four levels of the two main effects, corresponding to the four conditions in total. The results of this permutation-based mass univariate cluster test were then followed up with a linear mixed model (with the modelling steps as described previously).

4.3 Results

4.3.1 Results of the behavioural data analysis

Regarding naming accuracies (see Table 4.1), a binomial generalised mixed model (with logit as the link function) showed neither a semantic relatedness effect (β = 0.239, SE = 0.202, 95%CI = [0.157, 0.635], z = 1.183, p = 0.237), nor a dominant classifier (C) congruency effect (β = 0.097, SE = 0.159, 95%CI = [-0.215, 0.410], z =

0.610, p = 0.542) nor an effect of classifier probability distribution similarity (β= 0.019, SE = 0.132, 95% CI = [-0.239, 0.277], z = 0.143, p = 0.886).

Table 4.1: Detailed information on the best-fitting model for naming accuracies

Formula: Naming accuracies $\sim$ Number of strokes + Frequency of distractor + Semantic relatedness (S+ vs. S-) + Length of distractor in characters+ Dominant classifier congruency (C+ vs. C-) + Similarity (D+ vs. D-) between classifier probability distributions + (1 | subject) + (1 | target)

Fixed effects	Estimate	95% CI [low, high]	z-value	Pr(> $\|z\|$)
(Intercept)	4.285	[2.733, 5.837]	5.412	< 0.001
Number of strokes	-0.175	[-0.430, 0.079]	-1.349	0.177
Frequency of the distractor	-0.020	[-0.171, 0.210]	0.201	0.840
Length of Distractor	0.191	[-0.130, 0.513]	1.168	0.243
S-	0.239	[-0.157, 0.635]	1.183	0.237
C+	0.097	[-0.215, 0.410]	0.610	0.542
D-	0.019	[-0.239, 0.277]	0.143	0.886
S-:C+	-0.118	[-0.464, 0.228]	-0.668	0.504
Random effects				
σ^2	1.000			
τ_{item}	2.280			
$\tau_{\text{participant}}$	1.206			
N_{item}	38			
$N_{\text{participant}}$	34			

ICC	0.516
Observations	3,196
Marginal/Conditional R^2	0.011/ 0.520

As for naming latencies for correct responses (see Table 4.2 and Figure 4.4), a generalised mixed effects model with Gamma distribution and identity link showed that (1) the semantically unrelated conditions have significantly shorter naming latencies (β = 18.675, SE = 5.076, 95% CI = [-28.623, -8.727], z = 3.679, p = 0.002) than the related conditions; (2) the dominant classifier congruency failed to reach significance (β = 6.782, SE = 3.599, 95% CI = [-0.272, 13.835], z = 1.884, p = 0.060); (3) Similarly, the similar classifier probability distribution conditions showed no significant difference from the dissimilar classifier probability distribution conditions (β = 6.084, SE = 3.362, 95% CI = [0.506, 12.674], z = 1.809, p = 0.070).

Table 4.2: Detailed information on the best-fitting model for naming latencies

Formula: Naming latencies ~ Number of strokes + Frequency of distractor + Semantic relatedness (S+ vs. S-) + Length of distractor in characters + Dominant classifier congruency (C+ vs. C-) + Similarity (D+ vs. D-) between classifier probability distributions + (1 | subject) + (1 | target)

Fixed effects	Estimate	95% CI [low, high]	z-value	Pr(> $\|z\|$)
(Intercept)	880.212	[861.168, 899.255]	90.594	< 0.001 * **
Number of strokes	-0.735	[-7.759, 6.289]	-0.205	0.837
Frequency of the distractor	-4.646	[-8.747, -0.545]	-2.220	0.026*
Length of Distractor	2.497	[-5.189, 10.183]	0.637	0.524
S-	-18.675	[-28.623, -8.727]	-3.679	0.002 ***

C+	6.782	[-0.272, 13.835]	1.884	0.060
D-	6.084	[-0.506, 12.674]	1.809	0.070
S-:C+	-0.509	[-7.769, 6.752]	-0.137	0.891
Random effects				
σ^2	0.041			
τ_{item}	1257.580			
$\tau_{\mathrm{participant}}$	2823.471			
N_{item}	38			
$N_{\mathrm{participant}}$	34			
ICC	1.000			
Observations	2,985			
Marginal/Conditional R^2		0.073/ 1.000		

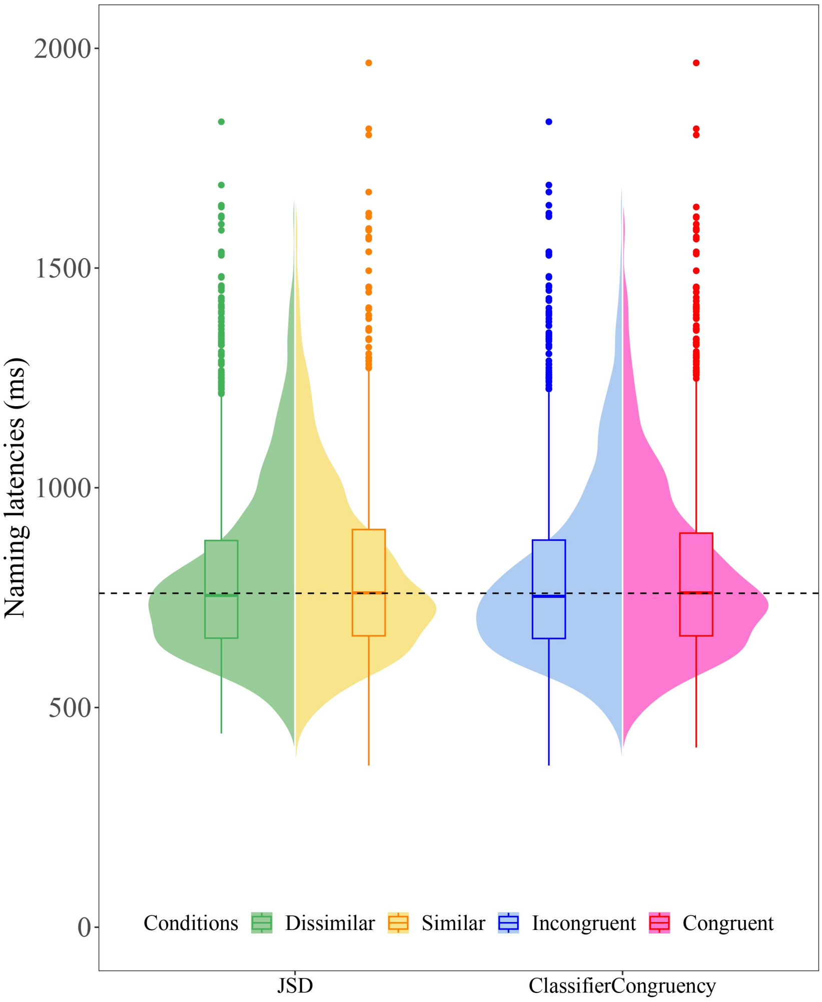


Figure 4.4: Naming latencies across the conditions of the Classifier Distribution Similarity and Dominant Classifier Congruency.

4.3.2 EEG data analysis

Results of planned analyses

Amplitude for a priori selected channels in the 275 - 575 ms time window (see Table 4.3, Figures 4.5 and 4.6 shows that (1) the semantically unrelated conditions have significantly more negative amplitudes (β = -0.262, SE = 0.013, df = 2964779.559, 95% CI = [0.287, -0.237], t = 20.463, $p < 0.001$) than the related conditions; (2) the dominant classifier-incongruent conditions have a significantly more negative amplitude relative to the congruent ones (β = -0.209, SE = 0.0122, df = 3700151.141, 95% CI = [0.186, -0.232], t = 17.816, $p < 0.001$); (3) the dissimilar classifier distributions have a significantly more positive amplitude relative to the similar classifier distribution (β = 0.122, SE = 0.011, df = 3690045.045, 95% CI = [0.101, 0.144], t = 11.159, $p < 0.001$). Last, the more negative dominant classifier effect was more negative for similar at the centro-parietal region (β = 0.040, SE = 0.014, df = 3706478.978, 95% CI = [0.068, 0.012], t = 2.817, $p = 0.005$). The larger classifier distribution similarity effect was more negative for similar at the centro-parietal region (β = 0.101, SE = 0.014, df = 3706478.980, 95%CI = [0.073, 0.129], t = 7.099, $p < 0.001$) than in other regions.

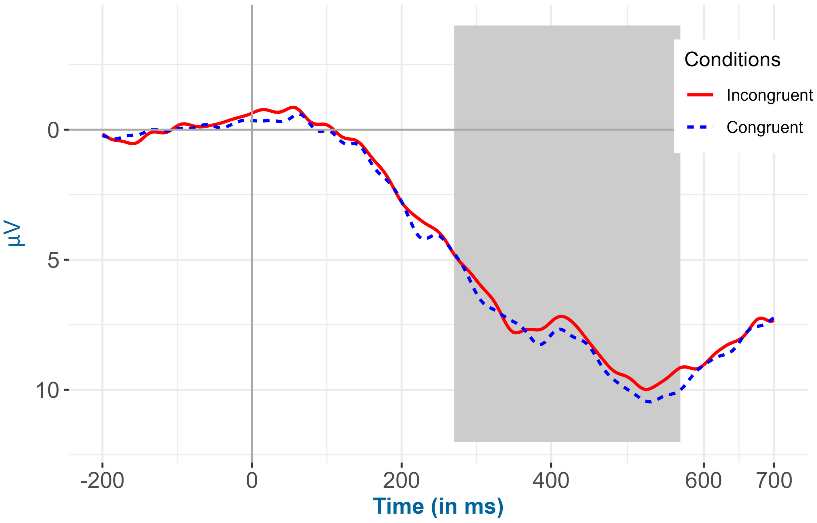


Figure 4.5: Mean amplitudes at CP1, CP5, P3, PO3, CP2, CP6, P4, and PO4 for dominant classifier congruent and incongruent conditions from – 200 ms to 700 ms (grey window is 275–575 ms).

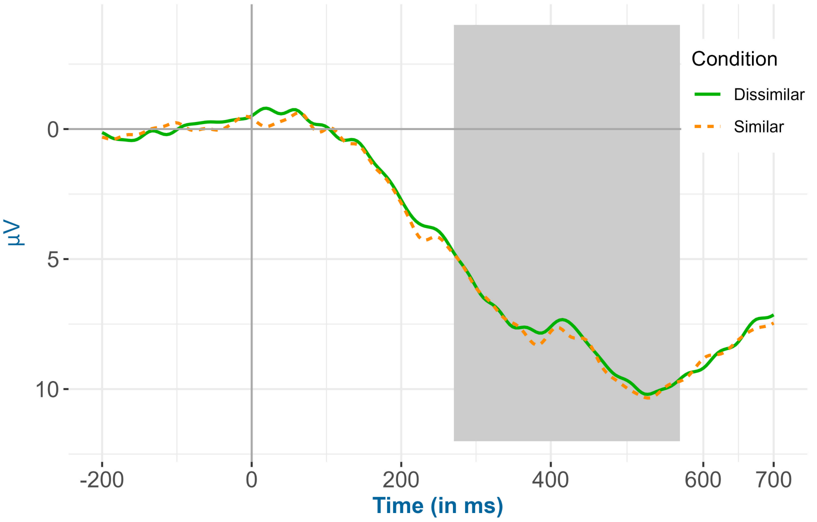


Figure 4.6: Mean amplitudes at CP2, CP6, P4, and PO4 for similar and dissimilar classifier probability distributions from – 200 ms to 700 ms (grey window is 275–575 ms).

Table 4.3: Details on the best-fitting model for electrophysiological response of F3, FC1, FC5, C3, CP1, CP5, P3, PO3, F4, FC2, FC6, C4, CP2, CP6, P4, and PO4 in the time window of 275-575 ms.

Formula: Voltage $\sim$ Time + Frequency of distractor + Number of strokes + Semantic relatedness (S + vs. S-) + Length of distractor in characters + Elec (centro-parietal vs. frontocentral) + Dominant classifier congruency (C + vs. C-) + Similarity (D- vs. D+) between classifier probability distribution + Semantic relatedness:Dominant classifier congruency + Classifier Congruency: Elec + Similarity between classifier probability distribution:Elec + (1|Subject) + (1|Target)

Fixed effects	Estimate	95% CI [low, high]	z-value	Pr($>|z|$)
(Intercept)	9.326	[7.593, 11.060]	10.682	< 0.001 ∗∗∗
Time	1.852	[1.840 1.864]	305.587	< 0.001 ∗∗∗
Frequency of the distractor	-0.150	[-0.161, -0.139]	-26.641	< 0.001 ∗∗∗
Number of strokes	0.179	[0.162, 0.196]	20.820	< 0.001 ∗∗∗
S-	-0.242	[-0.267, -0.217]	-18.807	< 0.001 ∗∗∗
Length of Distractor	-0.181	[-0.202, -0.161]	-17.448	< 0.001 ∗∗∗
Elec: frontocentral	-2.542	[-2.567, -2.517]	-196.916	< 0.001 ∗∗∗
C+	0.216	[0.193, 0.239]	18.410	< 0.001 ∗∗∗
D-	0.138	[0.116, 0.159]	15.531	< 0.001 ∗∗∗
S-:D-	-0.148	[-0.168, -0.129]	15.103	< 0.001 ∗∗∗

Elec frontocentral: C+	-0.040	[-0.068, -0.01]	2.818	0.005 ∗∗∗
Elec frontocentral: D-	0.101	[0.073, 0.129]	7.100	< 0.001 ∗∗∗
Random effects				
σ^2	136.113			
τ_{item}	1.097			
$\tau_{\text{participant}}$	24.878			
N_{item}	38			
$N_{\text{participant}}$	34			
ICC	0.160			
Observations	3,706,560			
Marginal/Conditional R^2		0.186/ 0.031		

Results of exploratory permutation-based TFCE analyses

A mass univariate cluster permutation test using a linear mixed model Amplitude $\sim$ congruency of semantic category + Number of strokes + Frequency of distractor + Conditions (the combined four levels of the two main effects) + (1 | Subject) + (1 | Target) and TFCE to control the Type-I error at 5% was performed (see Figure 4.7). The cluster is in the centro-parietal area and occurs between 275–425 ms post-stimulus onset. Based on the permutation test results, the following channels were selected for the cluster: CP5, P7, P3, PO3, and O1.

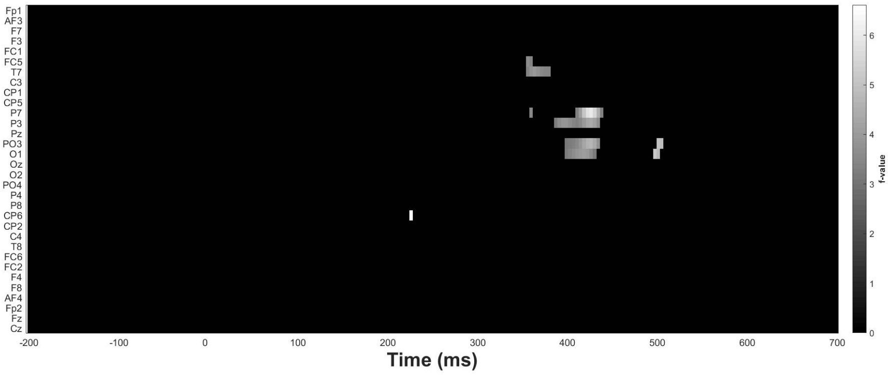


Figure 4.7: Result of permutation-based TFCE analyses.

For the cluster highlighted in Figure 4.7, the mixed model revealed similar results in the exploratory permutation-based TFCE analyses as in the analyses reported above. That is (see Table 4.4, Figures 4.8 and 4.9), (1) semantically unrelated conditions have less negative amplitudes ($\beta = 0.157$, SE = 0.025, df = 162746.157, 95% CI = [0.108, 0.206], t = 6.278, $p < 0.001$) relative to the semantically related conditions; (2) dominant classifier- incongruent conditions display significantly more negative amplitudes compared to dominant classifier-congruent conditions ($\beta = 0.210$, SE = 0.018, df = 579156.103, 95% CI = [0.174, -0.246], t = 11.473, p <0.001); (3) the dissimilar classifier probability distribution condition has a significantly more positive amplitude relative to the similar classifier probability distribution ($\beta = 0.142$, SE = 0.016, df = 527393.715, 95% CI = [0.110, 0.174], t = 8.730, $p < 0.001$).

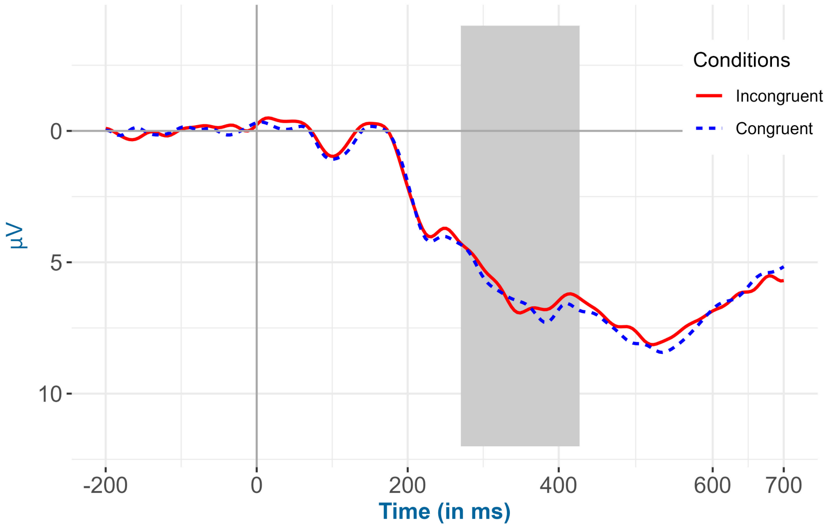


Figure 4.8: Mean amplitudes at CP5, P7, P3, PO3, and O1 for congruent vs. incongruent dominant classifier conditions from – 200 ms to 700 ms (grey window is 275 - 425 ms).

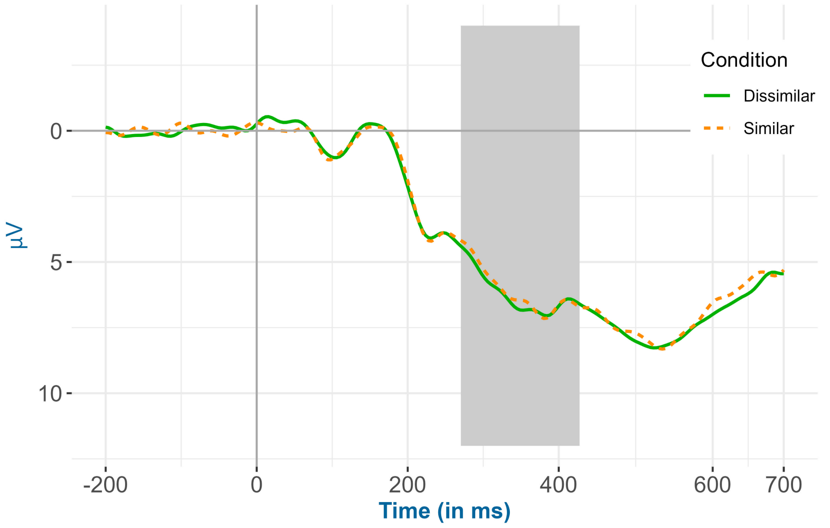


Figure 4.9: Mean amplitudes at CP5, P7, P3, PO3, and O1 for similar vs. dissimilar classifier probability distributions from – 200 ms to 700 ms (grey window is 275 - 425 ms).

Table 4.4: Detailed information on the best-fitting model for electrophysiological response of CP5, P7, P3, PO3, and O1 at the time-window of 275 - 425 ms.

Formula: Voltage $\sim$ Time + Frequency of distractor + Number of strokes + Semantic relatedness (S + vs. S-) + Length of distractor in characters + Dominant classifier congruency (C + vs. C-) + Similarity (D- vs. D+) between classifier probability distribution + Semantic relatedness:Dominant classifier congruency + (1|subject) + (1|target)

Fixed effects	Estimate	95% CI [low, high]	z-value	Pr(> $\lvert z \rvert$)
(Intercept)	7.039	[5.635, 8.443]	9.950	$< 0.001 ***$
Time	0.655	[0.631, 0.678]	55.002	$< 0.001 ***$
Frequency of the distractor	-0.129	[-0.150, -0.107]	-11.659	$< 0.001 ***$
Number of strokes	0.068	[0.035, 0.101]	4.066	$< 0.001 ***$
S-	0.160	[0.110, 0.209]	6.352	$< 0.001 ***$
Length of Distractor	-0.022	[0.062, 0.018]	-1.096	0.273
C+	0.211	[0.175, 0.247]	11.509	$< 0.001 ***$
D-	0.144	[0.112, 0.176]	8.797	$< 0.001 ***$
S-:C+	-0.117	[-0.154, -0.079]	6.057	$< 0.001 ***$

Random effects

σ^2	86.223

τ_{item}	0.514
$\tau_{\text{participant}}$	16.343
N_{item}	38
$N_{\text{participant}}$	34
ICC	0.164
Observations	608,850
Marginal/Conditional R^2	0.005/ 0.168

4.4 Discussion

To summarise, we found that producing an intended Mandarin Chinese noun in the presence of distractor nouns that have dissimilar classifier probability distributions results in a more positive ERP amplitude relative to distractor nouns having a similar classifier probability distribution in the time window 275-425/575 ms after stimulus onset. However, we did not observe any effects of classifier probability distribution similarity at the behavioural level. Regarding dominant classifiers, we found that producing an intended Mandarin Chinese noun in the presence of dominant classifier- incongruent distractor nouns results in more negative amplitudes of the electrophysiological response in the same time window 275–425/575 ms compared to dominant classifier-congruent distractor nouns. At the behavioural level, we did not observe any dominant classifier effects either. Regarding semantic relatedness, at the behavioural level, we observed longer naming latencies for related conditions compared to unrelated conditions, replicating the well-established semantic interference effect (Wang et al., 2019; see Bürki et al., 2020 for a meta-study). At the electrophysiological level, we observed a more negative effect for semantically unrelated vs. related conditions in planned analyses in the time window between 275-575 ms at the centro-parietal region (CP1, CP5, P3, PO3, CP2, CP6, P4, and PO4). However, we observed a more positive effect for semantically unrelated vs. related conditions in permutation-based TFCE analyses in the time window between 275–425 ms at the left centro-parietal-occipital region (CP5, P7, P3, PO3, and O1).

The successful replication effect of longer naming latencies for semantically related vs. unrelated conditions indicates that our experimental design could detect expected effects. The electrophysiological effect we observed under such conditions in the planned analysis is opposite to the findings of the permutation- based TFCE analysis but in line with previous studies (Blackford et al., 2012; Costa et al., 2009; Greenham et al., 2000; Wang et al., 2019). We argue that this contradiction results from the permutation test being conducted over the combined four levels comprising both dominant classifier congruency and classifier distribution similarity probability. This combination of levels during the TFCE analysis results in a mismatch between the criteria used for determining the spatiotemporal windows and those used for analyzing the semantic relatedness effect. Thus, the permutation-based TFCE analysis yields different results regarding semantic relatedness from that of the planned analysis and literature (Blackford et al., 2012; Costa et al., 2009; Greenham et al., 2000; Wang et al., 2019) - which could explain the contradictory findings.

Regarding the dominant classifier congruency effect, we observed more negative amplitudes for the electrophysiological responses of the incongruent dominant classifiers compared to congruent dominant classifiers around 400 ms post stimulus onset with a maximum at centro-parietal regions, in line with the classic N400 effect. This N400 effect, combined with the absence of any behavioural effect for the dominant classifier incongruent vs. congruent conditions, is in line with our prediction and with previous studies that classifiers are activated but not selected (Wang et al., 2019; Wang et al., 2024). Therefore, we will not discuss the dominant classifier effect and its implications in more detail here. Instead, we will focus our discussion in the following section on the classifier probability distribution effect and its implications for language production models, specifically the model developed by Levelt and colleagues.

4.4.1 Classifier probability distribution effect

To the best of our knowledge, the present study is the first to investigate whether models of language production can be extended to situations where multiple options for lexico-syntactic features are grammatically allowed. Current models of language production have been constructed based on previous findings in Indo- European languages where single grammatically correct options are available. For instance, La Heij et al. (1998) and Starreveld & La Heij (2004) conducted behavioural studies on grammatical gender in Dutch and concluded that the lexico-syntactic features are not selected in bare noun naming because no behavioural gender congruency effects were found in PWI tasks. More recent studies wherein electrophysiological results were combined with behavioural observations concluded that, although not ultimately selected, the dominant classifier is still activated (Wang et al., 2019; Wang et al., 2024).

In the current study, we investigated the role of the similarity between the classifier probability distributions of target and distractor words whilst controlling for dominant classifier congruency effects and semantic relatedness. We did not observe any effect of the similarity between classifier probability distributions at the behavioural level. Even though this absence of the classifier probability distribution effect at the behavioural level might be due to a Type II error (i.e. false negative), it could also imply that none of the classifiers that make up the classifier probability distribution are selected. This line of thinking is consistent with the reasoning employed in existing literature (La Heij et al., 1998; Starreveld & La Heij, 2004; Wang et al., 2019; Wang et al., 2024). Regarding the electrophysiological responses, we observed more positive amplitudes for dissimilar vs. similar classifier probability distribution conditions.

We attributed this electrophysiological effect to the P600-like effect based on the following reasons. First, we manipulated a lexico-syntactic feature (similarity of classifier probability distributions), which is known to elicit P600 effects in prior research (Wang et al., 2024) and has been suggested to be elicited for theoretical reasons (Hagoort & Brown, 2000; Popov et al., 2020; Wang et al., 2019). Second, the peaks of ERP waveforms for each condition were positive-going and located around 550 ms post-stimulus onset, which aligns with the typical P600 component time window (although the observed differences span the 275-575 ms time window) (Kappenman & Luck, 2012). Finally, the effect was maximal at centro-parietal electrodes, which has also been observed previously (Wang et al., 2024) and is consistent with the typical P600 effect (Hagoort et al., 1993; Osterhout & Holcomb, 1992).

Assuming that the P600 effect is syntactically driven, we speculate that when producing a given bare noun (1) multiple compatible classifiers are activated with the degree of activation being determined by their compatibility with the given noun even though none of them is finally selected (La Heij et al., 1998; Levelt, 1999; Levelt et al., 1999; Starreveld & La Heij, 2004) and (2) that the activation of these classifiers is lexico-syntactically driven.

4.4.2 Implications for Levelt's language production model

Based on the findings and associated interpretations of the current study, we propose an extension of Levelt's model of language production to allow it to accommodate lexico-syntactic features that have multiple grammatically correct options (see Figure 4.10). Note that we attributed the observed effects to lexical level processing as Mandarin Chinese classifiers belong to the family of lexico-syntactic features, and we controlled for semantic relatedness in data analysis.

To illustrate, producing a particular noun such as, e.g. "arm", the noun will first have to be activated at the conceptual level and then lexicalised at the lexical level before finally being articulated. The lexicalised item representing the noun, located at the lemma-level, has a unidirectional link from the lexicalised item to its corresponding classifier probability distribution. Through this link, automatic activation of multiple compatible classifiers occurs when producing a bare noun, albeit to differing degrees. The degree of ac-

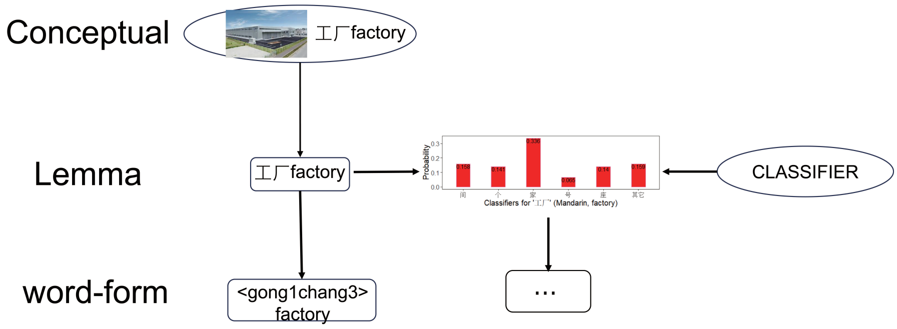


Figure 4.10: Illustration of proposed modified version of Levelt's model of language production incorporating the case where lexico-syntactic features have multiple grammatically correct options, represented herein by the Mandarin Chinese classifiers as an example. CLASSIFIER here is a placeholder (see also Wang et al., 2019)

tivation for an individual classifier could depend on its compatibility with the given noun. For instance, according to the classifier probability distribution associated with 工厂(/gong1chang3/, "factory", the normalised co-occurrence (i.e. compatibility) for the classifiers 条(/tiao2/) and 只(/zhi1/(are 0.246 and 0.579, respectively. Thus, we speculate that when producing 工厂(/gong1chang3/, "factory") the classifiers 条(/tiao2/) and 只(/zhi1/) are activated with their degrees of activation being determined by their compatibility with the given noun 工厂(/gong1chang3/, "factory"). Classifiers which are fully incompatible with a given noun we consider not to be activated.

However, since classifiers are not needed for the task of producing a bare noun, none of the classifiers will be finally selected. In case classifiers are needed, speakers will select the classifier that is either (1) most strongly activated (Zhan & Levy, 2018), or (2) closest to their intended meaning. For example, 张(/zhang1/) is used for 邮票(/you2piao4/, "stamp") when emphasising the stamp as a long and thin entity, while 枚(/mei2/) is used when emphasising the stamp as a small item (Wu & Bodomo, 2009).

As such, this extension of Levelt's model of language production can accommodate both the case where lexico-syntactic features have multiple grammatically correct options and the case where lexico-syntactic features have single grammatically correct options.

4.5 Conclusion

In conclusion, the current study investigated the encoding of Mandarin Chinese classifiers during speech production, where multiple lexico-syntactic features are possible for a given noun. Our results indicate that, when producing a bare noun, multiple compatible classifiers are activated. These results extend our understanding of the underlying mechanism of speech production when there is more than one grammatically correct option for a lexico-syntactic feature for a given noun.

4.6 Acknowledgements

This work was partially funded by the China Scholarship Council (CSC) no. 202006070023. We are grateful to all members of the LUCL Experimental Linguistics Lab, our anonymous reviewers, and all our participants for their contributions.

4.7 Disclosure statement

No potential conflict of interest was reported by the author(s).

4.8 Funding

This work was supported by China Sponsorship Council: [grant number 202006070023]. Niels O. Schiller is supported by CityUHK startup grant no. 9380177 from City University of Hong Kong.

Appendix 4

Appendix 4A. Stimuli used in the experiment

Target	Distractor	Distribution Similarity	Classifier Congruency
蚂蚁 /ma2yi/ ant	手套 /shou3tao4/,glove	Dissimilar	Incongruent
	蟒蛇 /mang3she2/,python	Dissimilar	Incongruent
	鸟/niao3/,bird	Similar	Congruent
	蜗牛/wo1niu2/,snail	Similar	Congruent
手臂 /shou3bi4/ arm	飞虫 /fei1chong2/, bug	Dissimilar	Incongruent
	工厂 /gong1chang3/, factory	Dissimilar	Incongruent
	脚 /jiao3/, foot	Similar	Incongruent
	鳄鱼 /e4yu2/, crocodile	Similar	Incongruent
凉鞋 /liang2xie2/ sandal	骆驼 /luo4tuo2/, camel	Dissimilar	Congruent
	风车 /feng1che1/, windmill	Dissimilar	Incongruent
	鞋子 /xie2zi/, shoe	Similar	Congruent

	拖鞋 /tuo1xie2/, slipper	Similar	Incongruent
冰鞋 /bing1xie2/ skate	梨 /li2/, pear	Dissimilar	Congruent
	矛 /mao2/, spear	Dissimilar	Incongruent
	碗 /wan3/, bowl	Similar	Congruent
	袋鼠 /dai4shu3/, kangaroo	Similar	Incongruent
熊 /xiong2/ bear	黄蜂 /huang2feng1/, wasp	Dissimilar	Congruent
	猿 /yuan2/, ape	Dissimilar	Incongruent
	企鹅 /qi3e2/, penguin	Similar	Congruent
海狸 /hai3li2/ beaver	鸵鸟 /tuo2niao3/, ostrich	Dissimilar	Congruent
	奶牛 /nai3niu2/, cow	Dissimilar	Incongruent
	仓鼠 /cang1shu3/, hamster	Similar	Congruent
麻袋 /ma2dai4/ sack	电视 /dian4shi4/, television	Dissimilar	Congruent
	长笛 /chang2di2/, flute	Dissimilar	Incongruent
	滑梯 /hua2ti1/, slide	Similar	Congruent
脸 /lian3/ face	水壶 /shui3hu2/, kettle	Dissimilar	Congruent
	贝雷帽 /bei4lei2mao4/, beret	Dissimilar	Incongruent
	桌子 /zhuo1zi/, desk	Similar	Congruent
心 /xin1/ heart	螺丝钉 /luo2si1ding1/, screw	Dissimilar	Congruent
	橙子 /cheng2zi/ orange	Dissimilar	Incongruent
	星星 /xing1xing1/, star	Similar	Congruent

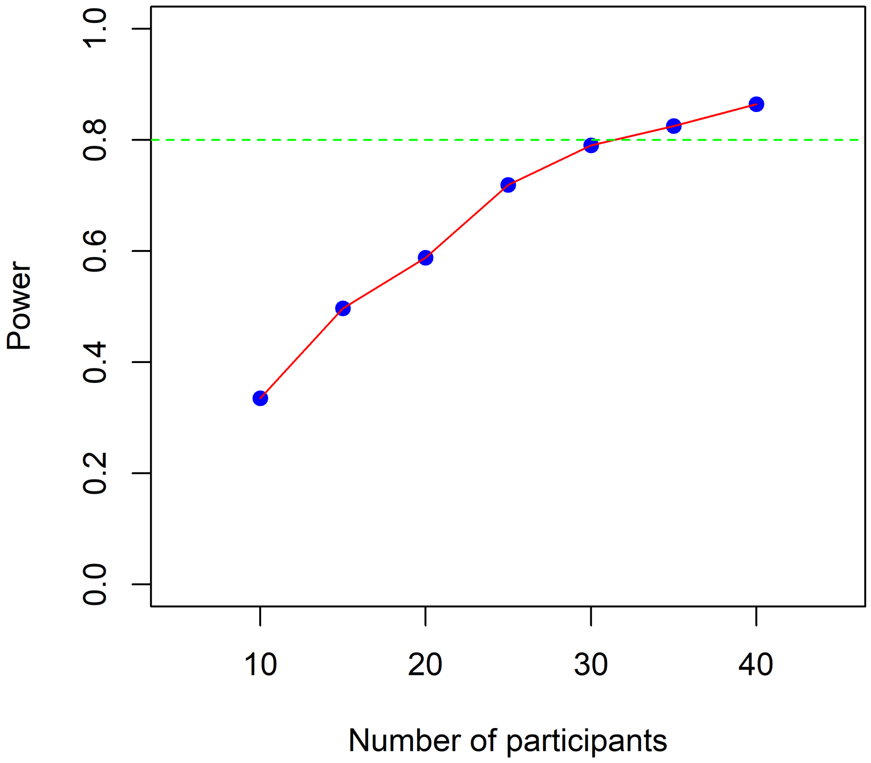
腿 /tui3/ leg	藤蔓 /teng2man4/, vine	Dissimilar	Congruent
	背心 /bei4xin/, vest	Dissimilar	Incongruent
	河 /he2/, river	Similar	Congruent
尾巴 /wei3ba1/ tail	面包 /mian4bao1/, bread	Dissimilar	Congruent
	鹰 /ying1/, eagle	Dissimilar	Incongruent
	长裤 /chang2ku4/, trousers	Similar	Congruent
腰带 /yao1dai4/ belt	瓢虫 /piao2chong2/, ladybug	Dissimilar	Congruent
	火鸡 /huo3ji1/, turkey	Dissimilar	Incongruent
	舌头 /she2tou4/, tongue	Similar	Congruent
比基尼 /bi3ji1ni2/ bikini	苹果 /ping2guo3/, apple	Dissimilar	Congruent
	大象 /da4xiang4/, elephant	Dissimilar	Incongruent
	盘子 /pan2zi/, plate	Similar	Congruent
短袜 /duan3wa4/ sock	大象 /da4xiang4/, elephant	Dissimilar	Congruent
	马 /ma3/, horse	Dissimilar	Incongruent
	青蛙 /qing1wa1/, frog	Similar	Congruent
尺子 /chi3zi/ ruler	火炬 /huo3ju4/, torch	Dissimilar	Congruent
	骰子 /shai1zi4/, dice	Dissimilar	Incongruent
	菜刀 /cai4dao1/, chopper	Similar	Congruent
萝卜 /luo2bo4/ carrot	冰淇淋 /bing1ji1ling2/, ice cream	Dissimilar	Congruent
	报纸 /bao4zhi3/, newspaper	Dissimilar	Incongruent

	螺栓 /luo2shuan1/, bolt	Similar	Congruent
汉堡 /han4bao3/ hamburger	马甲 /ma3jia2/, waistcoat	Dissimilar	Congruent
	潜艇 /qian2ting3/, submarine	Dissimilar	Incongruent
	背包 /bei1bao1/, backpack	Similar	Congruent
抽屉 /chou1ti2/ drawer	大篷车 /da4peng2che1/, caravan	Dissimilar	Congruent
	马车 /ma3che1/, carriage	Dissimilar	Incongruent
	月亮 /yue4liang4/, moon	Similar	Congruent
衣柜 /yi1gui4/ wardrobe	锅 /guo1/, pot	Dissimilar	Congruent
	天鹅 /tian2e2/, swan	Dissimilar	Incongruent
	圆 /yuan2/, circle	Similar	Congruent
项链 /xiang4lian4/ necklace	鲨鱼 /sha1yu2/, shark	Dissimilar	Congruent
	斑马 /ban1ma3/, zebra	Dissimilar	Incongruent
	领带 /ling3dai4/, tie	Similar	Congruent
手指 /shou3zhi3/ finger	芦苇 /lu2wei3/, reed	Dissimilar	Congruent
	灯塔 /deng1ta3/, lighthouse	Dissimilar	Incongruent
	拇指 /mu2zhi3/, thumb	Similar	Incongruent
手 /shou3/ hand	葡萄 /pu2tao2/, grape	Dissimilar	Incongruent
	苍蝇 /cang1ying2/, fly	Similar	Congruent
城堡 /cheng2bao3/ castle	头盔 /tou2kui1/, helmet	Dissimilar	Incongruent
	小丘 /xiao3qiu1/, hill	Similar	Congruent

披风 /pi1feng1/ cloak	冰屋 /bing1wu1/, igloo	Dissimilar	Incongruent
	毛衣 /mao2yi1/, sweater	Similar	Congruent
裤子 /ku4zi/ pants	燕麦 /yan4man4/, oats	Dissimilar	Incongruent
	头巾 /tou2jing1/, turban	Similar	Congruent
衬衫 /chen4shan1/ shirt	香蕉 /xiang1jiao1/, banana	Dissimilar	Incongruent
	夹克 /jia2ke4/, jacket	Similar	Congruent
钢笔 /gang1bi3/ fountain pen	卡车 /ka3che1/, truck	Dissimilar	Incongruent
	笔 /bi3/, pen	Similar	Congruent
铅笔 /qian1bi3/ pencil	刷子 /shua1zi/, brush	Dissimilar	Incongruent
	箭 /jian4/, arrow	Similar	Congruent
鹰嘴豆 /ying1zui3dou4/ chickpeas	监狱 /jian1yu4/, prison	Dissimilar	Incongruent
	蝴蝶 /hu2die2/, butterfly	Similar	Congruent
玉米 /yu4mi3/ corn	黄油 /huang2you2/, butter	Dissimilar	Congruent
	火车 /huo3che1/, train	Dissimilar	Incongruent
蜡烛 /la4zhu2/ candle	蛋糕 /dan4gao1 / cake	Dissimilar	Incongruent
	烟斗 /yan1dou2/, pipe	Similar	Congruent
长椅 /chang2yi3/ bench	耳环 /er2huan2/, earring	Dissimilar	Incongruent
	凳子 /deng4zi/, stool	Similar	Congruent
闹钟 /nao4zhong1/ alarm clock	刺猬 /ci4wei3/, hedgehog	Dissimilar	Incongruent

	马克杯 /ma3ke4bei1/, mug	Similar	Congruent
厨师 /chu2shi1/ chef	竹林 /zhu3lin2/, forest	Dissimilar	Incongruent
	医生 /yi1sheng1/, doctor	Similar	Congruent
蜜蜂 /mi4feng1/ bee	鲤鱼 /li3yu2/, carp	Dissimilar	Incongruent
	松鼠 /song1shu3/, squirrel	Similar	Congruent
耳朵 /er2duo1/ ear	手提箱 /shou3ti2xiang1/, suitcase	Dissimilar	Incongruent
	鸭子 /ya1zi/, duck	Similar	Congruent
眼睛 /yan3jing1/ eye	骨头 /gu3tou2/, bone	Dissimilar	Incongruent
	山羊 /shan1yang2/, goat	Similar	Congruent
戒指 /jie4zhi3/ ring	火箭 /huo3jian4/, rocket	Dissimilar	Congruent
	直升机 /zhi2sheng1ji1/, helicopter	Dissimilar	Incongruent
	手榴弹 /shou3liu2dan4/, grenade	Similar	Congruent

Appendix B. Power curve for determining the sample size of participants


Power
Number of participants