The activation and selection of lexico-syntactic features in speech production: behavioural and electrophysiological evidence from L1 and L2 speakers

Wang, S.

CHAPTER 4

Does native language affect grammatical gender processing? ERP signatures of lexico-syntactic feature similarity in non-native production: the case of English-Dutch and Chinese-Dutch speakers

Abstract
In this study, we studied whether or not gender processing in L2 speech production is affected by similarity in lexico-syntactic features in L1 and L2. We compared two groups: English learners of Dutch, whose L1 lacks a lexico-syntactic feature, and a group of Chinese learners of Dutch, whose L1 possesses classifiers, a lexico-syntactic feature similar to grammatical gender in L2. We manipulated gender congruency and semantic relatedness in a picture-naming task using the picture-word interference paradigm. Behaviorally, we observed the semantic interference effect but did not find a pronounced gender congruency effect in the English-Dutch group, which was nearly comparable to the Chinese-Dutch group, nor did we find a feature-similarity effect. At the neural level, we observed a similarity effect associated with semantic relatedness in N400 amplitudes. Unexpectedly, we did not find evidence for a similarity effect related to gender congruency in voltage amplitudes. Therefore, our results suggest a limited role of similarity in classifiers and gender in L2 gender processing. This study contributes to gaining more insights into L2 production in light of lexico-syntactic features.

Keywords
Speech production; Gender congruency effect; Semantic interference effect; Similarity; Event-related potentials

4.1. Introduction

In Chapters 2 and 3, we investigated the processing patterns of L2 Chinese classifiers in Dutch late learners of Chinese and the processing patterns of L2 Dutch grammatical gender in Chinese late learners of Dutch. We also compared them with L1 speakers of the languages involved in each chapter, which were L1 Chinese and L1 Dutch speakers. Through our exploration, we found some similarities in processing patterns for both classifiers and gender, in line with previous studies (e.g., Huang & Schiller, 2021; Wang et al., 2019). Given the similarity in classifier and gender processing, a question arises: in the same group of people, such as Chinese-Dutch speakers with classifiers in their L1 and gender in their L2, will they have an advantage in gender processing in L2 production over speakers without classifiers and gender in their L1? In this chapter, we expanded the previous findings and experimental settings to another language group: the English-Dutch group. Distinct from Chinese and Dutch, the English language does not grammatically employ nominal gender or other lexico-syntactic features (Tokowicz & MacWhinney, 2005). Therefore, English and Dutch can be considered as dissimilar in lexico-syntactic features compared to Chinese and Dutch.

Some studies have observed that speakers with similar gender systems in both their L1 and L2 are more likely to perform similarly to native speakers in the L2 compared to speakers with dissimilar gender systems (De Diego Balaguer et al., 2005; Dussias et al., 2013; Jeong et al., 2007; Sabourin et al., 2006; Sabourin, 2001; Schepens et al., 2013; Van Der Slik, 2010). For instance, Sabourin et al. (2006) and Sabourin (2001) found that in Dutch L2 gender processing in Dutch gender agreement and assignment tasks, speakers of German L1 with highly similar gender systems had the highest accuracy, followed by Romance L1 speakers with dissimilar gender systems, but English L1 speakers, who lack a gender system, had significantly worse accuracy.

A few electroencephalographic (EEG) and neuroimaging studies suggest that similar features in L2 and L1 exhibit similar neural patterns (Foucart & Frenck-Mestre, 2011; Jeong et al., 2007; Sabourin & Stowe, 2008; Tokowicz & MacWhinney, 2005). For instance, Foucart and Frenck-

Mestre (2011) found that high-proficiency German-French speakers showed a native-like P600 effect when processing gender agreement violations, but only when the rules for determiner and noun were similar in their L1 and L2, while there was no effect when the rules were dissimilar. Sabourin and Stowe (2008) observed that German-Dutch speakers, whose gender systems in both languages were highly similar, showed a native-like P600 effect elicited by gender agreement violations in L2, while Romance-Dutch speakers, whose gender systems in both languages were dissimilar, did not display this effect. These studies suggest that gender processing in L2 can be impacted by the degree of similarity across languages.

Moreover, individual variations such as proficiency levels can also affect L2 gender processing. For instance, Dussias et al. (2013) observed that high-proficiency late English-Spanish speakers performed in a more native-like manner, processing gender information in their L2 quickly and efficiently, outperforming low-proficiency English-Spanish speakers. This finding from English learners aligned with EEG results, in which Dowens et al. (2010) observed that high-proficiency late English-Spanish learners showed similar P600 effects in L2 gender processing to native speakers. Their findings suggest that late language learners with sufficiently high proficiency can perform in a native-like manner in L2 gender processing, no matter the dissimilarity of lexico-syntactic features between L1 and L2.

The question is to what extent language similarity affects L2 processing. Given previous studies have shown that the similarity of gender systems affects gender processing in L2, can classifiers fall into this framework? For instance, can classifiers in L1 affect gender processing in L2, either facilitating or interfering with it, compared to the case in which there is no gender or classifier in L1? To answer this question, we investigated gender processing in L2 Dutch by English-Dutch learners, whose L1 lacks gender and classifiers. We compared their performance with that of the Chinese-Dutch learners with intermediate proficiency as presented in Chapter 3. Moreover, to align with the average proficiency level of Chinese-Dutch speakers and considering that there might be more influence of similarity on L2 learners with lower than advanced proficiency, the English-Dutch speakers in the current study have an average intermediate proficiency. Given that most studies have focused on high-proficiency L2

speakers, with limited research involving intermediate proficiency levels, exploring intermediate learners helps fill a research gap.

One model that may related to the current study is the *Language Distance Hypothesis* (LDH) (Zawiszewski & Laka, 2020). This theoretical account predicts that highly similar features in L1 and L2 would lead to similar behavioral and neural patterns, while dissimilar features in L1 and L2 would result in different patterns. This could be reflected in higher accuracy, shorter response times, and relatively more native-like ERP effects, especially for the gender congruency effect, but not for the semantic interference effect. This difference is expected between the group with similar features in L1 and L2 compared to the group with dissimilar features.

4.1.1. The current study

In line with Chapter 3, the current study first examined how gender congruency and semantic relatedness influence L2 Dutch NP production by English-Dutch speakers. Subsequently, we compared English-Dutch with Chinese-Dutch speakers to examine the impact of similarity on L2 production based on whether L1 and L2 shared lexico-syntactic features. We built upon the findings from previous chapters of this thesis, which observed the similarities in the processing of classifier and gender features in L1 and L2 production. We explored the potential advantages of L2 gender processing arising from the similarity in L1 and L2 features through a comparison between the English-Dutch and Chinese-Dutch groups, based on the LDH account.

We applied the methodologies employed in Chapter 3, conducting an identical picture-naming task using the PWI paradigm combined with electroencephalography (EEG) and event-related potentials (ERP) to test English-Dutch speakers. The first effect we focus on is the *gender congruency effect* (e.g., Schriefers, 1993; for reviews, see Schiller, 2013 as well as Wang & Schiller, 2019; for a meta-analysis see Bürki et al., 2023). This effect represents the competitive selection process of the gender feature and is reflected as faster naming speed when target pictures and distractor words share the same gender value (e.g., the target "*de* hond_{com}" [the dog] - the distractor word "*de* $\text{tafel}_{\text{com}}$" [the table] in Dutch), compared to when the pictures and distractor words have different gender values (e.g., the target "*de* hond_{com}" [the dog] - the distractor word "*het* boek_{neu}" [the book] in

Dutch). The second effect we expected is the *semantic interference effect* that mirrors the competitive selection of lexical entries (e.g., Schriefers, Meyer, & Levelt, 1990; for a meta-analysis, see Bürki et al., 2020). The effect is reflected as faster naming when the pictures and distractor words belong to different semantic categories (e.g., to-be-named picture of a dog with the distractor words "table"), compared to when the pictures and distractor words belong to the same semantic category (e.g., the picture of a dog with the distractor words "cat"). Considering that the semantic interference effect is typically associated with lexical access according to the LRM model (Levelt et al., 1999), we included this effect as a criterion for assessing word processing by L2 speakers. The absence of this effect may suggest less efficient access to lexical information and/or the activation and competition of corresponding lemmas.

For the English-Dutch group, based on previous studies (e.g., Levelt et al., 1999; Schiller & Caramazza, 2003, 2006) and results from Chapter 3, we predicted faster naming in the semantically unrelated condition compared to the related condition, as well as faster naming in the gender-congruent condition compared to the incongruent condition, demonstrating the semantic interference effect and the gender congruency effect. Regarding the comparison between both groups, our predictions were as follows: First, we expected potential advantages for the Chinese-Dutch group with similar features over the English-Dutch group in terms of naming accuracy and latencies for gender congruency rather than semantic relatedness. Second, we expected distinct gender congruency effects between the two groups. Third, we expected that the observed ERP patterns, modulated by gender congruency, would differ between the two groups, with smaller ERP patterns for the English-Dutch group compared to the Chinese-Dutch group. The ERP patterns possibly manifest as either an N400 effect (Paolieri et al., 2020; Wicha et al., 2003) or a P600 effect (Barber & Carreiras, 2005; Caffarra & Barber, 2015; Hagoort, 2003; Hagoort & Brown, 1999; Molinaro et al., 2011).

4.2. Methods

4.2.1. Participants

For the English-Dutch group, we recruited twenty (15 females; *M* = 29 years of age, *SD* = 8.8) healthy, right-handed native English speakers with an A2 and higher proficiency level of Dutch in accordance with the CEFR (Council of Europe, 2001) from Leiden University and Delft University in the Netherlands. Participants reported no psychological or language disorders or visual and hearing impairments. Before taking part in the experiment, all of them signed informed consent forms in compliance with the Ethics Code for linguistic research in the Faculty of Humanities at Leiden University.

English-Dutch speakers' linguistic profiles and experience with Dutch were assessed through an adapted version of the language history questionnaire (LHQ; Li et al., 2020). At the time of testing, all participants were residing in the Netherlands. For the eighteen participants included in the analysis (see Section 3.1), two participants reported Dutch as the second language, nine participants reported as the third, and seven participants as the fourth. The mean AoA of the Dutch language was M_{AoA} = 23.85 years of age (*SD* = 8.04). The mean year of Dutch language use was *M* = 3.93 (*SD* = 5.14). On a scale from zero to seven (with seven being maximally proficient), participants self-reported their current proficiency in Dutch-speaking as *M* = 2.89 (*SD* = 1.60), in reading as *M* = 3.83 (*SD* = 1.54), in listening as *M* = 3.50 (*SD* = 1.72) and in writing as *M* = 2.78 (*SD* = 1.31).

We noticed that three participants reported using Dutch over ten years (10, 12, and 17 years of use, respectively), which was longer than other participants (with *M* = 3.93, *SD* = 5.14). These three participants had a mean AoA of the Dutch language of M_{AoA} = 26.33 years of age (*SD* = 4.16). On a scale from zero to seven (seven being maximally proficient), they reported a speaking proficiency of *M* = 4.67 (*SD* = 2.08), a reading proficiency of *M* = 5.33 (*SD* = 1.53), a listening proficiency of *M* = 5.00 (*SD* = 2.00), and a writing proficiency of *M* = 4.00 (*SD* = 2.00). Their Dutch proficiency was supposed to be higher than other participants based on their self-reported information.

Moreover, eighteen participants reported learning one or two other languages prior to acquiring Dutch (M_{AoA} = 9.08 years of age, *SD* = 8.34), with fourteen participants reported acquiring Romance languages (i.e., French, Spanish, or Romanian) prior to Dutch (M_{AoA} = 9.50 years of age, *SD* = 6.17). Four participants had learned Germanic languages (i.e., German or Afrikaans) prior to Dutch (M_{AoA} = 9.60 years of age, *SD* = 10.85). Additionally, five participants each had acquired one of the following languages: Indonesian, Filipino, Serbian, Russian, or Mandarin, prior to Dutch (M_{AoA} = 7.86 years of age, *SD* = 11.23). See Appendix 4.A.1 for an overview of the languages acquired by participants. Among them, five participants reported acquiring their second language (Afrikaans, Filipino, German, Romanian, or Russian) almost simultaneously with English, with a mean AoA of *M* = 1 year of age (*SD* = 1.1). These five simultaneous speakers reported a speaking proficiency of *M* = 6.40 (*SD* = 0.55), a reading proficiency of *M* = 6.40 (*SD* = 0.55), a listening proficiency of *M* = 7.00 (*SD* = 0), and a writing proficiency of *M* = 5.60 (*SD* = 1.14) in their second language. It is important to note that all five participants reported a higher self-reported proficiency in English compared to the simultaneously acquired languages. Therefore, English is the dominant language among these participants.

To assess whether the Dutch proficiency of L1 English participants and L1 Chinese participants from Chapter 3 are matched, we conducted the same LexTale-Dutch task (Lemhöfer & Broersma, 2012) as in Chapter 3. The mean LexTale-Dutch score in the English-Dutch group was *M* = 60.97 (*SD* = 9.93). For the Chinese-Dutch group, the mean LexTale-Dutch score was *M* = 57.83 (*SD* = 13.14) (see Chapter 3). An ANOVA analysis showed no significant difference in LexTale scores between the two groups with F (1, 35) = 0.668, p = 0.419. See Section 4.2.2.2 for proficiency tests and Section 4.3.3 for more details about LexTale-Dutch scores in the English-Dutch group.

4.2.2. Tasks and materials

Before the experiment, participants performed a picture-naming task and a LexTale-Dutch test (Lemhöfer & Broersma, 2012) in the lab. The picture-naming task was programmed in E-prime 3 (Psychology Software Tools, Pittsburgh, PA). We measured EEG during the picture-naming task.

Picture-naming task

The tasks and materials were the same as those used in Chapter 3. More specifically, the task adopted a 2 × 2 factorial within-subject design using the PWI paradigm. We controlled gender congruency (congruent versus incongruent) and semantic relatedness (related versus unrelated) as the two factors, resulting in four experimental conditions: gender congruent and semantically related (G+S+), gender incongruent and semantically related (G-S+), gender congruent and semantically unrelated (G+S-), and gender incongruent and semantically unrelated (G-S-). Participants saw 26 target pictures four times, and each was assigned a condition. Pictures were selected from the MultiPic database (Duñabeitia et al., 2018). Half of the pictures had common gender according to the gender value of the pictures' names. The other half of the pictures had neuter gender. Each picture was paired with four distractor words. We selected distractor words primarily from "Nederlands in gang" (Berna et al., 2017) and "Nederlands in actie" (Berna et al., 2012), which are Dutch textbooks used in Dutch language courses at Leiden University, in order to ensure that the stimuli are familiar to language learners. Each pair of target pictures and distractor words was assigned to one of four experimental conditions, resulting in a total of 104 target-distractor word pairs, which were presented in a pseudo-random order. The pseudo-random lists were generated using the Windows program Mix (Van Casteren & Davis, 2006) in the following rule: a minimum of ten trials separated two identical target pictures, allowing for the consecutive appearance of either three identical conditions with different pictures or three different pictures with identical determiners.

Proficiency tests

To keep matters consistent with Chapter 3, we conducted two proficiency tests to evaluate the proficiency of English-Dutch speakers. One was the gender knowledge test, which consisted of 20 sentences with a blank space for the determiner before the noun (e.g., "__ tafel" [the table]). The test was generated using Qualtrics software (Qualtrics, 2021). The other test was an online LexTale-Dutch task (Lemhöfer & Broersma, 2012). This task was used to assess participants' vocabulary size in Dutch and supplement their self-reported proficiency levels. The task score was included in the statistical analysis.

4.2.3. EEG recording

EEG data were collected using a standard thirty-two Ag/AgCI electrode 10/20 montage with the BrainVision Recorder software by BrainProducts. The vertical electrooculogram (*VEOG*) was recorded using two external flat electrodes placed below and above the participant's left eye. The horizontal electrooculogram (*HEOG*) was recorded using two electrodes placed at the outer canthus of the eyes. Two flat electrodes were placed on the mastoid positions behind the participant's ears for subsequent data re-referencing.

4.2.4. Procedure

Participants were provided with an information sheet and signed an informed consent form prior to participation, in line with the ethics guidelines at the Faculty of Humanities at Leiden University. They received reimbursement after their participation. During the experiment, participants were seated in front of a computer screen in an experiment booth.

The picture-naming task was divided into a familiarization session, a practice session and an experimental session. In the familiarization session, participants were exposed to all target pictures with Dutch names underneath and were instructed to get familiar with the pictures. In the practice session, participants were presented with pictures again and were instructed to overtly name each picture in Dutch using an NP construction with the correct determiner and noun (e.g., "de hond" [the dog]). Each picture was superimposed by a string of two X's, i.e., "XX". During this session, the experimenter provided oral feedback when the participant incorrectly produced either the determiner or the noun, or both. In the experimental session, participants were instructed to name the picture as fast and accurately as possible and to reduce movement. Pictures were superimposed by Dutch distractor words. We recorded participants' behavioral responses and EEG during the experimental session. A typical trial was initiated with the display of a fixation cross "+" in the center of the screen for 300 ms, followed by a blank screen for 250 ms, and the display of the picture for 3,000 ms in the center of the screen, and a blank screen for 300 ms before the start of the next trial. Each pair of picture-distractor words was shown only once during the experimental session.

Subsequently, participants performed the original web-based LexTale-Dutch test in the lab. Participants were instructed to make a lexical decision on whether or not the string on the screen was a Dutch word. The accuracy and scores were generated online. The result was included as a covariate in the analysis of the picture-naming task.

4.3. Results

4.3.1. Behavioral data exclusion

One participant was excluded for failing to complete the picture-naming task. Another participant was excluded to keep consistency with the EEG data analysis (see Section 3.4 for details). We included eighteen data sets in the analysis. For the analysis of naming latencies, we removed 21.58% from all data points because of (a) participants responding incorrectly or exceeding the time limit, resulting in incomplete responses (20.19%); (b) outliers (i.e., naming latencies exceeding 3 *SD*s of a participant's mean reaction time; 1.39%).

4.3.2. Behavioral data analysis

We first used Praat (Boersma & Weenink, 2021) to extract *naming latencies* and calculated *accuracy* for the picture-naming task. Then, we applied a mixed-effect model approach using the *lme4* package to model *naming accuracy* and *naming latencies* in RStudio (Version 2023.03.0.). For *naming accuracy*, we used linear mixed-effect models (GLMMs) with a binomial distribution using the *glmer()* function to model for all trials. For *naming latencies*, we performed a generalized linear mixed effect model (GLMM) using the *glmer()* function with a gamma distribution and the identity link function to model the positively skewed naming latencies for correct trials (Matuschek et al., 2017). Our fixed-effects structure consisted of *gender congruency* (congruent vs. incongruent) and *semantic relatedness* (related vs. unrelated). Moreover, we included several covariates, such as *determiner (de* vs. *het)*, *LexTale-Dutch score*, *age of acquisition*, *year of use*, *learning Romance languages before Dutch*, *learning Asian languages before Dutch*, and *the order of Dutch acquisition*. To conduct a combination analysis for both English-Dutch and Chinese-Dutch groups, we included *group* (Chinese-Dutch vs. English-Dutch) as a covariate to interact with both

gender congruency and *semantic relatedness* in the model. Our random-effects structure consists of *participant* and *item* (i.e., the target) as well as random slopes for the main manipulation. We used treatment coding as our contrast. The default reference levels were set as follows: *gender congruency* was set to *congruent*, *semantic relatedness* to *related*, *accuracy* to *correct*, and *group* to *Chinese-Dutch group*. We discarded corrections between random slopes and interactions between fixed effects in the case of non-convergence or singular fit. We compared models with different structures using the *anova()* function to establish the best-fitting model.

4.3.3. Behavioral data results

We first calculated the mean naming latencies and accuracy for the remaining participants in the English-Dutch group. See Table 1 for descriptives of *naming latencies* (ms) and *accuracy* (% correct) by condition, as well as the descriptives for the Chinese-Dutch group (from Chapter 3). We plotted naming latencies for each group in Figure 1.

The mean LexTale-Dutch score in the English-Dutch group was M = 60.97 (SD = 9.93). Scores ranged from 40 and 91.95, with 100 being the possible maximum score. According to Lemhöfer and Broersmas's (2012) classification scheme, vocabulary scores below 60 were related to "B1" and lower levels of proficiency, and 60 – 80 were related to "B2" levels. Therefore, the average proficiency level of our speakers was categorized as the "upper-intermediate (B2)" level. However, there are differences within participants. Specifically, nine speakers were classified as the "lower-intermediate (B1 and lower)" level, eleven speakers could be classified as the "upper intermediate (B2)" level, and one speaker was classified as "upper & lower advanced (C1 & C2)". The average score of the gender knowledge test for the English-Dutch group was M = 83.06 (SD = 8.02). The maximal score was 100. Two participants who did not pass the inclusion criterion of 75% correction were excluded before the experiment.

Accuracy

For the English-Dutch group, the final model for *naming accuracy* included main effects for *semantic relatedness* and *gender congruency*, and random effects for *participant* and *item*. The interaction of two fixed effects did not improve the model fit with $\chi 2(1) = 0.148$, $p = 0.7$ and was removed

from the model. Covariables *AoA* ($\chi2(1) = 0.905$, $p = 0.341$), *self-rating score* ($\chi2(1) = 2.278$, $p = 0.131$), *year of use* ($\chi2(1) = 3.628$, $p = 0.057$), *learning Romance languages before Dutch* ($\chi2(1) = 0.014$, $p = 0.907$), *learning Asian languages before Dutch* ($\chi2(1) = 0.019$, $p = 0.890$), *learning German before Dutch* ($\chi2(1) = 3.440$, $p = 0.064$), and *the order of Dutch acquisition* ($\chi2(1) = 2.372$, $p = 0.124$) did not significantly improve the model fit and were removed. *Determiner* did not improve the model fit with $\chi2(1) = 2.155$, $p = 0.142$. A main effect of *LexTale-Dutch score* was observed with *Log-Odds* = -0.051, *SE* = 0.021, *95% CI* [-0.092, -0.010], $z = -2.412$, $p = 0.016$, suggesting the impact of proficiency on accuracy. Neither *gender congruency* nor *semantic relatedness* demonstrated a significant effect on *naming accuracy*. See Appendix 4.B.1 for model parameters.

Combining the data from the English-Dutch group and the Chinese-Dutch group (data from Chapter 3) for the accuracy analysis, despite a descriptive trend of higher accuracy in the Chinese-Dutch group compared to the English-Dutch group (see Table 1), the group difference in accuracy was not statistically significant, with *Log-Odds* = 0.521, *SE* = 0.317, *95% CI* [-0.101, -1.143], $z = 1.642$, $p = 0.101$. Neither *gender congruency* nor *semantic relatedness* significantly influenced both groups' accuracy. The 3-way interaction among *group*, *gender congruency*, and *semantic relatedness* ($\chi2(1) = 0.877$, $p = 0.928$), the interaction between *group* and either *gender congruency* or *semantic relatedness* ($\chi2(1) = 0.711$, $p = 0.701$), as well as the interaction between *gender congruency* and *semantic relatedness* ($\chi2(1) = 0.035$, $p = 0.852$) did not improve the model fit and be excluded. See Appendix 4.B.2 for model parameters.

Naming latencies

For the English-Dutch group, the final model for *naming latencies* included main effects for *semantic relatedness* and *gender congruency* and random effects for *participant* and *item*. The interaction of *gender congruency* and *semantic relatedness* caused non-convergence and was removed from the model. Covariables *LexTale-Dutch score* and *age of acquisition* resulted in non-convergence or singular fit and were dropped from the model fitting procedure. *Year of use* ($\chi2(1) = 1.933$, $p = 0.164$), *self-rating score* ($\chi2(1) = 2.233$, $p = 0.135$), *learning Romance languages before Dutch* ($\chi2(1) = 0.029$, $p = 0.864$), *learning Asian languages before*

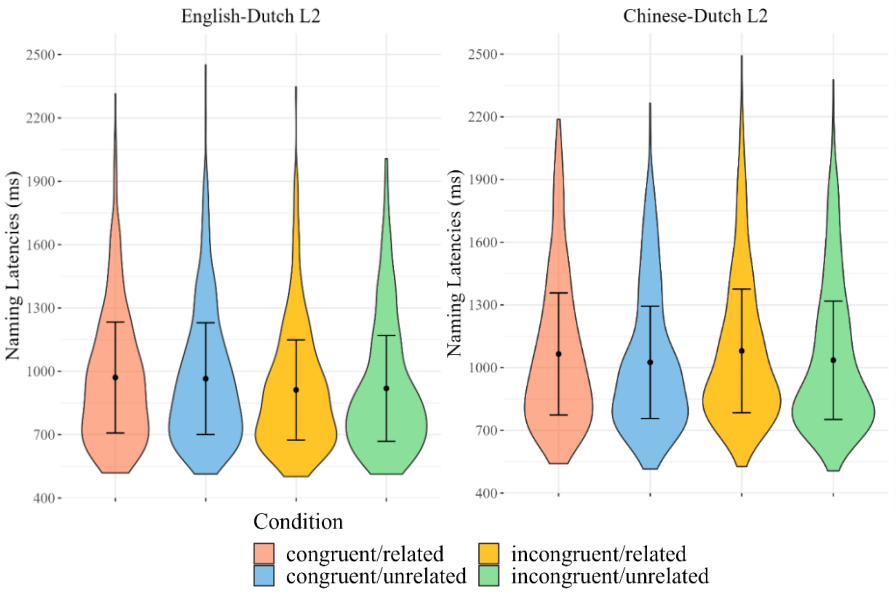
Dutch ($\chi 2(1) = 0.804$, $p = 0.370$), *learning German before Dutch* ($\chi 2(1) = 0.304$, $p = 0.582$), and *the order of Dutch acquisition* ($\chi 2(1) = 0.470$, $p = 0.493$). *Determiner* did not improve the model fit with $\chi 2(1) = 0.119$, $p = 0.731$ and was also removed. English-Dutch speakers were significantly faster at naming in the semantically unrelated condition compared to the related condition with $\beta = -0.040$, $SE = 0.010$, *95% CI* [-0.059, -0.021], $t = -4.134$, $p < 0.001$. However, *gender congruency* did not significantly influence *naming latencies*, with $\beta = -0.002$, $SE = 0.010$, *95% CI* [-0.021, 0.016], $t = -0.257$, $p = 0.797$. Naming latencies were comparable between gender-congruent and incongruent conditions, in line with descriptive results (see Table 4.1). See Appendix 4.C.1 for model parameters, and Figure 4.1 (left panel) for a visualization of the naming latencies in the English-Dutch group.

Comparing the English-Dutch and Chinese-Dutch groups, there was no significant group difference in naming latencies, with $\beta = -0.049$, $SE = 0.084$, *95% CI* [-0.214, 0.115], $t = -0.590$, $p = 0.555$. The 3-way interaction among *group*, *gender congruency* and *semantic relatedness* did not converge and was removed. The interaction between *group* and *gender congruency* or *semantic relatedness* ($\chi 2(1) = 0.58$, $p = 0.758$), as well as the interaction between *gender congruency* and *semantic relatedness* ($\chi 2(1) = 0.68$, $p = 0.41$) did not significantly improve the model fit and were removed from the modeling procedure. Both groups' naming was significantly faster in the *semantically unrelated* condition compared to the *related* condition with $\beta = -0.042$, $SE = 0.008$, *95% CI* [-0.057, 0.027], $t = -5.474$, $p < 0.001$. In contrast, no significant effect of *gender congruency* was observed in the combination analysis. See Appendix 4.C.2 for model parameters. These behavioral results suggest no significant evidence that feature-similarity (*group*) modulated *accuracy* and *naming latencies* between the two groups.

Table 4.1: *Mean naming latencies (ms) and mean accuracy (%) for each condition in the picture-naming task for the English-Dutch L2 group (n = 18), and for the Chinese-Dutch L2 group (n = 19, from Chapter 3, Table 3.1).*

		Latencies (ms)		Accuracy (%)	
Group	Condition	Mean	SD	Mean	SD
English-Dutch L2	congruent-related (G+S+)	971	328	78.9	15.5
	incongruent-related (G-S+)	965	331	77.4	13.7
	congruent-unrelated (G+S-)	912	296	81.0	15.4
	incongruent-unrelated (G-S-)	919	313	81.0	16.7
	Gender				
	congruent	941	314	80.0	15.2
	incongruent	942	322	79.2	14.9
	estimated difference	**1**		0.8	
	Semantic				
	related	968	329	78.1	14.4
	unrelated	916	305	81.0	15.9
	estimated difference	**52**		2.9	
Chinese-Dutch L2	congruent-related (G+S+)	1065	365	85.9	12.8
	incongruent-related (G-S+)	1080	370	89.9	11.6
	congruent-unrelated (G+S-)	1026	336	88.4	12.2
	incongruent-unrelated (G-S-)	1036	354	86.4	11.5
	Gender				
	congruent	1045	351	87.1	11.8
	incongruent	1057	362	88.1	10.8
	estimated difference	**12**		1	
	Semantic				
	related	1073	367	87.9	11.8
	unrelated	1031	345	87.4	11.1
	estimated difference	**42**		0.5	

Figure 4.1: *Mean naming latencies by four conditions with 95% confidence intervals in the English-Dutch group (left panel) and the Chinese-Dutch group (right panel, from Chapter 3, Figure 3.1).*



4.3.4. EEG data exclusion

One English-Dutch participant who was excluded from the behavioral analysis was also removed from the EEG analysis. The EEG data were included based on the following reasons: first, the participant produced a correct NP; second, the segments contained valid trials and did not contain artifacts. Participants with more than 60% valid trials could be included in the analysis. Subsequently, we excluded one more participant due to exceeding these criteria.

4.3.5. EEG data pre-processing

We performed pre-processing to separate the signal from artifacts using Brain Vision Analyzer 2.2 (Brain Products GmbH, 2013). The pre-processing procedure included the following steps: re-referencing to the average of the left and right mastoid, filtering between 0.1 Hz and 30 Hz with a high- and low-pass filter, linear derivation for the two *HEOG* and two *VEOG* electrodes to form two combined channels for horizontal and vertical

eye movements, respectively, semi-automatic ocular correction using *VEOG* and *HEOG* parameters, and finally, artifact rejection. We then computed segments around the picture onset triggers for each participant from -200 ms prior to onset to 1,200 ms after picture onset. Subsequently, we performed baseline correction on the 200 ms signal prior to picture onset. We computed a segment for each condition, resulting in four segments for each participant per trial.

4.3.6. EEG data analysis

In order to explore the locus of the effect of gender congruency and semantic relatedness on voltage amplitudes we performed permutation tests using the *permute* package (Voeten, 2019) in RStudio (Rstudio Team, 2023) to calculate the F-values across 26 electrodes (excluding *Fp1, Fp2, AF3, AF4, T7,* and *T8*). We first explored the effect of semantic relatedness, then explored the effect of gender congruency. The outcomes of permutation tests suggested a potential semantic effect in centro-parietal regions in the time window between approximately 250 ms and 450 ms after target onset. However, the effect of gender congruency did not appear evident in the outcome plot, making it difficult to clearly identify the specific channels and time windows where the gender effect might be present. Consequently, we selected ten channels in bilateral centro-parietal regions as our ROI: *C3, C4, Cz, CP1, CP2, CP5, CP6, P3, P4,* and *Pz,* along with the time window of 250 ms to 450 ms, based on the potential semantic effect observed in permutation analysis in the English-Dutch group. See Appendix 4.D.1 for the results of permutation tests. We then created grand-averaged plots, see Figure 4.2.

In the next step, we combined the EEG data from both English-Dutch and Chinese-Dutch groups into a single model to conduct a combination analysis. In the Chinese-Dutch group, we first observed an N400-like effect in five centro-parietal channels (*C3*, *C4*, *Cz*, *CP1*, and *CP2*) in the time window of 250 ms to 500 ms post-stimulus onset; meanwhile, a P600-like effect was observed in ten channels in broader centro-parietal regions (*CP1, CP2, P3, P4, Pz, PO3, PO4, O1, O2, and Oz*) in the time window of 500 ms to 800 ms post-stimulus onset (see Chapter 3).

Therefore, based on the outcome of the permutation analysis for the English-Dutch group and the findings from the Chinese-Dutch group, we first selected the time window shared by both groups: 250 ms to 450 ms post-stimulus onset, and five channels shared by both groups: *C3*, *C4*, *Cz*, *CP1*, and *CP2*. Then, as a complement, we explored the time window of 500 ms to 800 ms post-stimulus onset in the combination analysis because the Chinese-Dutch group observed a P600-like effect in this time window. We selected five channels shared by both groups, *CP1*, *CP2*, *P3*, *P4*, and *Pz*, for the P600-like effect analysis.

For the statistical analysis, we used the *lme4* package in RStudio based on the work by Frömer et al. (2018). First, we modeled a maximal LMM model using a normal distribution with *gender congruency* (congruent vs. incongruent) and *semantic relatedness* (related vs. unrelated) as fixed effects. The maximal model consists of the interaction between *gender congruency* and *semantic relatedness*, a covariate *hemisphere* (left vs. right vs. midline), random slopes for each *participant* and *item*, and by-*participant* random slopes for the interaction of *gender congruency* and *semantic relatedness*. Covariates *determiner (de* vs. *het)*, *LexTale-Dutch score*, *age of acquisition*, *years of usage*, *self-rating score*, *learning Romance languages before Dutch*, *learning Asian languages before Dutch*, *learning German before Dutch*, and the *order of Dutch acquisition* were included in the maximal model. In the combination analysis, we added *group* (Chinese-Dutch vs. English-Dutch) as a covariate and interacted *group* with *gender congruency* and *semantic relatedness*. We used treatment coding as our contrast. The default reference levels were *gender-congruent*, *semantically related* and *Chinese-Dutch group*. In the case of non-convergence, we removed covariates and removed the interaction of random effects, evaluating the significance of each by-*participant* random slope and removing the interaction of fixed effects. The model fitting procedure was similar to the behavioral analysis.

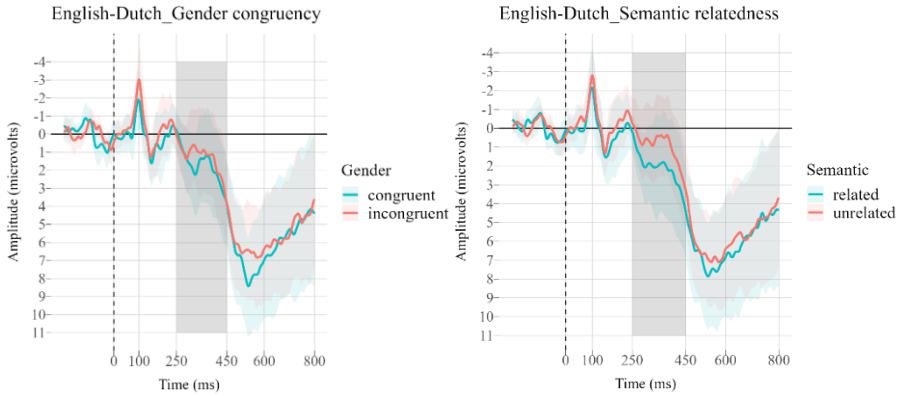
4.3.7. EEG data results

The English-Dutch group

Visual inspection of Figure 4.2 revealed an N1/P2 complex associated with early visual processing (Eulitz et al., 2000; Misra et al., 2012; Schendan & Kutas, 2003). Further, voltage amplitudes, which were modulated by

semantic relatedness, diverged around 250 ms, reached a small peak at 400 ms, and converged at around 450 ms post-stimulus onset.

Figure 4.2: *Mean voltage amplitudes by gender congruency (left panel) and by semantic relatedness (right panel) for centro-parietal channels C3, C4, Cz, CP1, CP2, CP5, CP6, P3, P4, and Pz for the English-Dutch group (n = 18). The N400 time window of interest from 250 ms to 450 ms is highlighted in grey. Note that negative voltage amplitudes are plotted upwards.*



The best-fitting model for the was as follows: voltage amplitudes ~ gender congruency (congruent vs. incongruent) + semantic relatedness (related vs. unrelated) + hemisphere (left vs. right vs. midline) + (semantic relatedness * gender congruency | participant) + (1| item). The interaction effect between *semantic relatedness* and *gender congruency* did not significantly improve the model fit with $\chi 2(1) = 0.0276$, $p = 0.868$ and was removed from the model. Covariates *LexTale-Dutch score* ($\chi 2(1) = 0.216$, $p = 0.642$), *age of acquisition* ($\chi 2(1) = 0.839$, $p = 0.360$), *years of usage* ($\chi 2(1) = 1.022$, $p = 0.312$), *learning German before Dutch* ($\chi 2(1) = 0.867$, $p = 0.710$), and *learning Asian languages before Dutch* ($\chi 2(1) = 1.528$, $p = 0.216$) did not improve the model fit and were also removed. The *order of Dutch acquisition*, *self-rating score*, and *learning Romance languages before Dutch* failed to converge. There was a main effect of *semantic relatedness* with $\beta = -1.138$, $SE = 0.502$, *95% CI* [-2.122, -0.153], $t = -2.265$, $p = 0.024$. The semantically unrelated condition elicited more negative amplitudes ($M_{S-} = 0.90$, $SD = 1.89$) compared to the related condition ($M_{S+} = 1.98$, $SD = 1.81$). However, there was no significant effect of *gender congruency*. See

Appendix 4.E.1 for model parameters, and Table 4.2 for average voltage amplitudes by condition. The results suggest that N400 voltage amplitudes were modulated by *semantic relatedness* rather than *gender congruency* in the English-Dutch group, which is consistent with the outputs of permutation tests and behavioral results.

Table 4.2: *Voltage amplitudes by condition for the N400 time window of 250 ms - 450 ms for channels C3, C4, Cz, CP1, CP2, CP5, CP6, P3, P4, and Pz for the English-Dutch group (n = 18).*

English-Dutch L2 group	**N400**	
Condition	Mean voltage(μV)	SD
congruent-related (G+S+)	2.14	1.87
incongruent-related (G-S+)	1.83	1.74
congruent-unrelated (G+S-)	0.94	1.96
incongruent-unrelated (G-S-)	0.87	1.82
Gender		
congruent	1.54	2.01
incongruent	1.35	1.84
estimated difference	**0.19**	
Semantic		
related	1.98	1.81
unrelated	0.90	1.89
estimated difference	**1.08**	

Combination analysis results

As described in Section 3.6, in order to explore the potential group-related effects on N400 and P600 voltage amplitudes, we combined the EEG data from both English-Dutch (n = 18) and Chinese-Dutch (n = 19) groups to perform LMM models.

For voltage amplitudes distributed across five channels (*C3*, *C4*, *Cz*, *CP1*, and *CP2*) in the time window of 250 ms to 450 ms post-stimulus onset, the best-fit model was as follows: voltage amplitudes ~ gender congruency (congruent vs. incongruent) * semantic relatedness (related vs. unrelated) * group (Chinese-Dutch vs. English-Dutch) + hemisphere (left vs. right vs. midline) + (gender congruency + semantic relatedness | participant) + (1| item). The interaction between *gender congruency* and *semantic relatedness*

significantly affected voltage amplitudes, with β = 0.538, *SE* = 0.073, *95% CI* [0.394, 0.682], t = 7.339, p < 0.001. Specifically, in the semantically related condition, voltage amplitudes were more negative when it was also in the gender-incongruent condition ($M_{G\text{-}S+}$ = 0.21, *SD* = 1.11) than in the congruent condition (M_{G+S+} = 0.31, *SD* = 1.10). In the semantically unrelated condition, voltage amplitudes were slightly more negative in the gender-incongruent condition ($M_{G\text{-}S\text{-}}$ = 0.07, *SD* = 1.09) than in the congruent condition ($M_{G+S\text{-}}$ = 0.08, *SD* = 1.08). There was a simple effect of *semantic relatedness* with β = -0.867, *SE* = 0.052, *95% CI* [-0.969, -0.765], t = -16.635, p < 0.001. Voltage amplitudes were more negative in the semantically unrelated condition ($M_{S\text{-}}$ = 0.08, *SD* = 1.09) than in the related condition (M_{S+} = 0.26, *SD* = 1.10).

Moreover, the interaction between *group* and *semantic relatedness* was significant with β = -0.446, *SE* = 0.077, *95% CI* [-0.596, -0.296], t = -5.83, p < 0.001. Specifically, in the semantically related condition, N400 voltage amplitudes for the English-Dutch group ($M_{S+:\,English}$ = 0.21, *SD* = 1.19) were more negative than for the Chinese-Dutch group ($M_{S+:\,Chinese}$ = 0.32, *SD* = 1.01). In the semantically unrelated condition, voltage amplitudes for the English-Dutch group ($M_{S-:\,English}$ = 0.05, *SD* = 1.15) were also more negative than for the Chinese-Dutch group ($M_{S-:\,Chinese}$ = 0.11, *SD* = 1.02). It appeared that the N400 effect, modulated by semantic relatedness, was bigger (more negative) in the English-Dutch group than in the Chinese-Dutch group. This result contrasts with our prediction, in which we expected no significant group differences in semantic processing. Moreover, there was no interaction effect between *group* and *gender congruency* (β = 0.149, *SE* = 0.670, *95% CI* [-1.165, 1.463], t = -0.222, p = 0.825), suggesting that voltage amplitudes showed a similar pattern in response to *gender congruency* for both groups. This result contradicts the predicted group difference in terms of the pattern of gender processing. See Appendix 4.F.1 for model parameters.

In Chapter 3, we observed a P600 effect in voltage amplitudes in centro-parietal regions in the time window between 500 ms and 800 ms post-stimulus. This effect was modulated by *semantic relatedness* and was specific to the Chinese-Dutch group. For the combination analysis, we combined data from both groups and modeled shared channels (*CP1, CP2,*

P3, *P4*, and *Pz*). The best-fit model was as follows: voltage amplitudes ~ gender congruency (congruent vs. incongruent) * group (Chinese-Dutch vs. English-Dutch) + semantic relatedness (related vs. unrelated) * group (Chinese-Dutch vs. English-Dutch) + hemisphere (left vs. right vs. midline) + (gender congruency + semantic relatedness | participant) + (1| item). The interaction between *gender congruency* and *semantic relatedness* did not significantly improve the model fit with $\chi 2(1) = 0.602$, $p = 0.74$ and was removed. Despite being included in the final model, there was no significant interaction between *group* and *gender congruency* or *semantic relatedness*, suggesting that voltage amplitudes did not significantly differ between the two groups. See Appendix 4.F.2 for model parameters.

4.4. Discussion

In this study, we examined gender congruency and semantic interference effects on speech production in English learners of Dutch. We employed a picture-naming task using the PWI paradigm and compared our findings with those of Chinese learners of Dutch in Chapter 3. We aimed to explore the potential influence of linguistic similarity at the level of lexico-syntactic features on L2 production, specifically investigating whether Dutch learners whose L1 possesses classifiers had a gender processing advantage compared to Dutch learners whose L1 lacks lexico-syntactic features, and whether or not the *Language Distance Hypothesis* (Zawiszewski & Laka, 2020) could be generalized to language pairs with classifier and gender features. We analyzed naming accuracy, naming latencies, and EEG data. We will first discuss the behavioral results.

In the English-Dutch group, participants were faster at naming in the semantically unrelated condition than in the related condition, displaying the typical semantic interference effect (e.g., Glaser & Düngelhoff, 1984; La Heij, 1988). This pattern is consistent with the findings from Chinese-Dutch participants, indicating that both Dutch L2 groups were sensitive to semantic information in their L2. This supports the selection-by-competition view, suggesting that activated lexical entries from the semantically related category compete with the target word and postpone its selection (e.g., Belke et al., 2005; Levelt et al., 1999). However, no gender congruency effect was

observed in the English-Dutch group, which showed nearly identical latencies between congruent and incongruent conditions, in contrast with the Chinese-Dutch group that displayed the trend of a gender congruency effect, even though the difference between conditions did not reach statistical significance.

Comparing both groups, we initially expected that the Chinese-Dutch group with similar lexico-syntactic features would show overall higher accuracy and faster production in terms of gender processing compared to the English-Dutch group. Specifically, we expected to observe different gender congruency effects between groups. However, despite the Chinese group displaying a trend of higher accuracy compared to the English group, and the English group showing a trend of faster naming compared to the Chinese group on average, we did not find a significant group effect in accuracy or naming latencies. This indicates that there is no significant overall difference between the groups. Moreover, regarding the group interaction with either gender congruency or semantic relatedness, we did not observe a significant interaction effect in accuracy or naming latencies, suggesting that the processing of gender and semantic features was comparable in both groups. The finding that no significant group difference based on gender congruency contrasts with our prediction but is consistent with previous studies (e.g., Costa et al., 2003).

In the ERP results, both the English-Dutch and Chinese-Dutch groups showed an effect of semantic relatedness. However, unlike the Chinese-Dutch group showed an interaction effect between gender congruency and semantic relatedness in N400 voltage amplitudes, the English-Dutch group did not display an interaction effect or a main effect of gender congruency in centro-parietal regions in N400 voltage amplitudes, as suggested by permutation tests and behavioral results. Comparing both groups, we found an interaction effect between similarity (*group*) and semantic relatedness in N400 voltage amplitudes, suggesting more negative amplitudes for the English-Dutch group than for the Chinese-Dutch group, both in semantically related and unrelated conditions. This finding was unpredicted, as we did not expect to find group differences in terms of semantic processing. However, we did not observe an interaction effect between similarity and gender congruency, contrasting with our hypothesis. Moreover, neither a main effect

of similarity nor an interaction effect was observed in P600-like voltage amplitudes, where we found a P600 effect modulated by semantic relatedness in the Chinese-Dutch group.

Taken together, we did not find evidence for distinct gender processing patterns between the English-Dutch group and the Chinese-Dutch group, which had similar overall proficiency levels. Our findings did not support our hypothesis that there were different gender congruency effects between the groups, specifically gender processing advantages for the Chinese-Dutch group over the English-Dutch group. Our results were in line with previous studies on cross-linguistic interaction at the gender level. For instance, Von Grebmer zu Wolfsthurn et al. (2022) found comparable ERP signatures between the gender-similar Italian-Spanish group and the gender-dissimilar German-Spanish group. Similarly, Costa et al. (2003) did not find similarity effects among Spanish-Catalan, Catalan-Spanish, Italian-French, and Croatian-Italian speakers. Therefore, the current study suggests that the feature-similarity in terms of lexico-syntactic features between L1 and L2 has a limited impact on gender processing in L2. These findings do not provide evidence for the generalization of the *Language Distance Hypothesis* (Zawiszewski & Laka, 2020) to language pairs with classifier and gender features.

As for the presentation of classifier and gender systems in the L1 and L2 lexicons of Chinese-Dutch speakers, our findings may align with the *gender-autonomous representation hypothesis* (Costa et al., 2003). This hypothesis suggests that gender systems are independent in L1 and L2 lexicons. It may can be extended to languages with classifiers, that is the classifier system in L1 and the gender system in L2 are autonomous without a clear correlation. One primary distinction between our participants and those in previous studies on cross-linguistic interference is the absence of a gender system in our participants' L1. Neither the English nor the Chinese lexicon include gender features. Consequently, based on the LRM model (Levelt et al., 1999), our results suggest that gender features in L2 Dutch theoretically receive activation exclusively from the lemma stratum in their Dutch lexicon. This means that the activation and selection of gender features in L2 Dutch are directly linked to the degree of development of their Dutch lexicon, which depends on proficiency. Therefore, proficiency could

be a possible explanation for the absence of a gender congruency effect in the English-Dutch group, as they were categorized as upper-intermediate levels in general, and their L2 Dutch lexicon was still in the process of development.

As a result, we hypothesize that as the proficiency in L2 Dutch increases in both English-Dutch and Chinese-Dutch groups, gender congruency effects in both groups will become more prominent. This hypothesis diverges from previous studies' conclusions that the gender congruency effect could increase with decreased L2 proficiency (Costa et al., 2003; Sá-Leite et al., 2020; Von Grebmer zu Wolfsthurn et al., 2021). The distinction can be attributed to whether or not participants' L1 have a gender system that requires the use of inhibitory control abilities to mitigate interference from the L1 on gender processing in the L2. This requirement does not apply to our participants because their L1 lack a gender system, and there is no prerequisite for such cross-linguistic interference at the gender level.

Another factor to consider is the potential influence of participants' other languages on the target language. Sixteen English-Dutch participants reported learning one or more languages before learning Dutch, such as French, Spanish and German, which have gender systems. Only two participants reported learning Dutch as their second language. In other words, the majority of English-Dutch participants had acquired knowledge of gender in other languages before learning Dutch. Although their self-reported proficiency in these languages was not high compared to their L1, their prior experience with gender language learning may have benefited their acquisition of Dutch. Therefore, variables arising from these language learning experiences should also be taken into account. In contrast, in the Chinese-Dutch group, eighteen participants reported English as their second acquired language, and sixteen participants reported Dutch as their third language (see Chapter 3, Appendix 3.A.1). In other words, in contrast to the English-Dutch participants, the majority of Chinese-Dutch participants did not acquire other gender languages before learning Dutch. Their learning experience in terms of gender feature knowledge appears to be less extensive compared to English-Dutch participants. To control for the potential effect of these variables, we included these collected language usage profiles in our

statistical models. However, most co-variables either led to the non-convergence of models or did not have a statistically significant impact on the results, as discussed in Sections 3.3.3 and 3.3.7. Nevertheless, we observed significant effects of AoA on the accuracy of the Chinese-Dutch group and LexTale score on the accuracy of the English-Dutch group. Accurately quantifying the impact resulting from these variations in learning experience is crucial for gaining comprehensive insights into L2 gender processing. Nevertheless, this task is beyond the scope of our study. Future research on non-native speakers can explore methods for further quantifying language profiles. Given the diversity among individuals within the same language group and the fact that most people are multilingual with varying language experience, achieving strict and uniform control over each participant's language background is exceedingly challenging. Developing a more precise quantification approach could enhance the reliability of results in non-native language studies.

4.5. Conclusion

In this study, we compared English-Dutch and Chinese-Dutch groups to explore whether or not native speakers with classifiers have an advantage in non-native gender language production. Our behavioral data showed the typical semantic interference effect in the English-Dutch group, but no evidence for a gender congruency effect, which was in line with the results of the Chinese-Dutch group. We did not find a feature-similarity effect in terms of accuracy and naming latencies, indicating that both groups had comparable performance. As for the EEG data, we observed an N400 effect modulated by semantic relatedness in the English-Dutch group. However, there was no significant ERP signature for a gender congruency effect. When comparing the two groups, we found a feature-similarity effect related to semantic relatedness on N400 amplitudes. Nevertheless, there was no feature-similarity effect associated with gender congruency on voltage amplitudes, suggesting similar gender processing patterns between the two groups. Therefore, our findings suggest a limited role of the similarity in lexico-syntactic features on L2 gender processing. Future studies could further investigate the relationship between proficiency, linguistic similarity, and L2 gender processing.

Declaration of Competing Interest

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

Credit author contribution statement

Shaoyu Wang: Conceptualization, Methodology, Programming, Data Acquisition, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review and Editing, Visualisation, Funding acquisition. **Niels O. Schiller**: Conceptualization, Methodology, Writing-Review and Editing, Supervision, Funding acquisition.

Acknowledgments

We are thankful to Aranka van Tol and Sarah von Grebmer zu Wolfsthurn for their comments on the EEG data analysis. We also thank Theres Grüter for her detailed comments on statistics and suggestions on results interpretation. We are grateful to our participants for their time in our study.

Appendix

4.A Linguistic profile: English-Dutch group

Table 4.A.1: *Overview of the native and non-native languages acquired by the current study participants (N = 18) according to the LHQ (Li et al., 2020).*

	L1	L2	L3	L4	Total
English	n = 18				**18**
Dutch		n = 2	n = 9	n = 7	**18**
French		n = 5	n = 3	n = 1	**9**
Spanish		n = 2	n = 2	n = 3	**7**
German		n = 3	n = 1		**4**
Romanian		n = 1	n = 1		**2**
Mandarin		n = 1		n = 1	**2**
Afrikaans		n = 1			**1**
Filipino		n = 1			**1**
Indonesian		n = 1			**1**
Russian		n = 1			**1**
Serbian			n = 1		**1**
Total	**18**	**18**	**17**	**12**	

4.B Model parameters: accuracy

Table 4.B.1: *Model of the best fit for accuracy in the English-Dutch L2 group, including estimated means, standard error, and confidence intervals (n = 18).*

Formula: naming accuracy ~ semantic (related vs. unrelated) + gender (congruent vs. incongruent) + LexTale score + (1| participant) + (1| item)					
Predictors	*Log-Odds*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	1.410	1.330	-1.197 – 4.016	1.060	0.289
Gender [incongruent]	0.057	0.127	-0.192 – 0.305	0.446	0.656
Semantic [unrelated]	-0.217	0.127	-0.466 – 0.032	-1.707	0.088
LexTALE	-0.051	0.021	-0.092 – -0.010	-2.412	**0.016**
Random Effects					
σ^2	3.29				
$\tau_{00\ Item}$	0.54				
$\tau_{00\ Subject}$	0.96				
ICC	0.31				
$N_{Subject}$	18				
N_{Item}	26				
Observations	1848				
Marginal R^2 / Conditional R^2	0.069 / 0.361				

Table 4.B.2: *Model of the best fit for combining accuracy in both English-Dutch and Chinese-Dutch groups, including estimated means, standard error, and confidence intervals (n = 37).*

Formula: accuracy ~ gender (congruent vs. incongruent) + semantic (related vs. unrelated) + group (Chinese-Dutch vs. English-Dutch) + (1| participant) + (1| item)					
Predictors	*Log-Odds*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	-2.240	0.277	-2.783 – -1.697	-8.092	<0.001
Gender [Incongruent]	0.012	0.094	-0.173 – 0.196	0.126	0.900
Semantic [unrelated]	-0.156	0.094	-0.341 – 0.029	-1.657	0.098
Group [English-DutchL2]	0.521	0.317	-0.101 – 1.143	1.642	0.101
Random Effects					
σ^2	3.29				
$\tau_{00\ Participant}$	0.83				
$\tau_{00\ Item}$	0.57				
ICC	0.30				
$N_{\ Participant}$	37				
$N_{\ Item}$	26				
Observations	3824				
Marginal R^2 / Conditional R^2	0.016 / 0.309				

4.C Model parameters: naming latencies

Table 4.C.1: *Model of the best fit for naming latencies in the English-Dutch L2 group, including estimated means, standard error, and confidence intervals (n = 18).*

Formula: naming latency ~ semantic (related vs. unrelated) + gender (congruent vs. incongruent) + (1| participant) + (1| item)

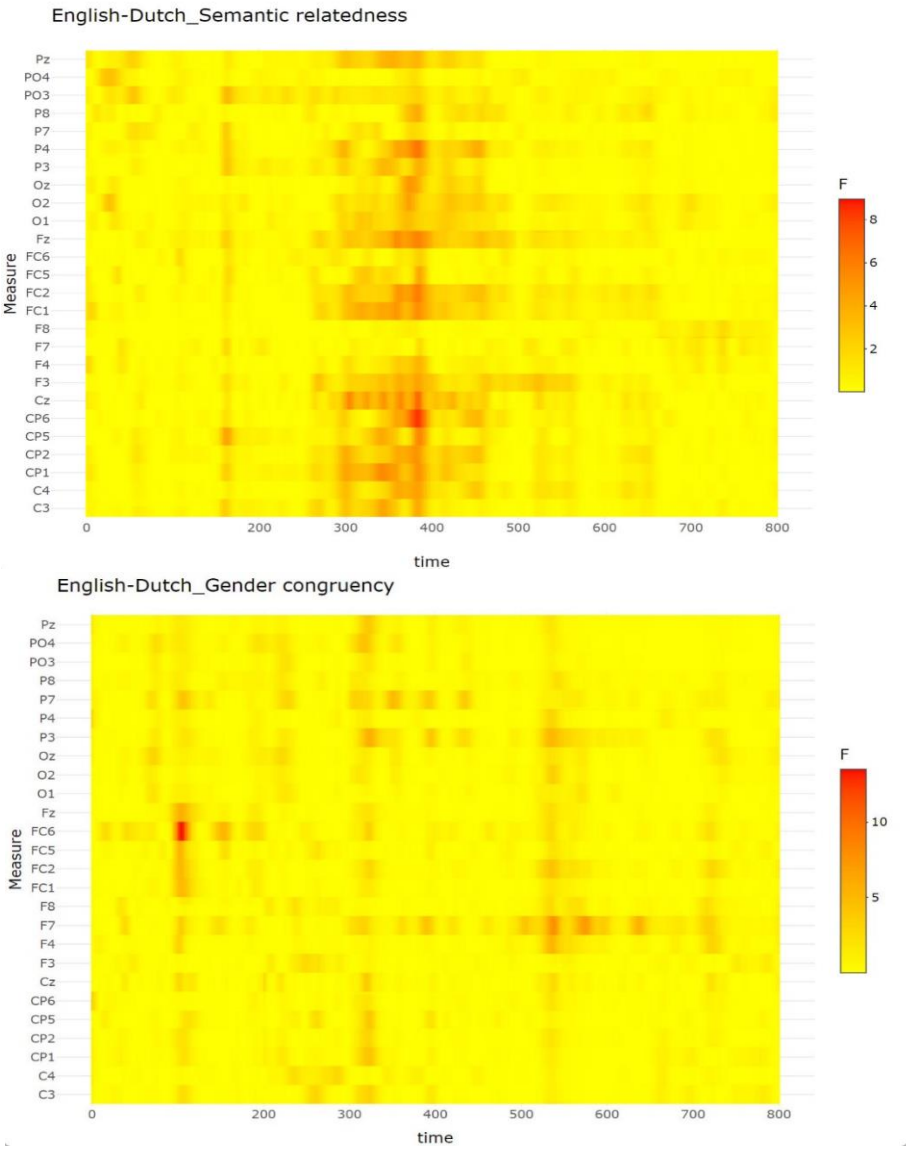
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	1.066	0.073	0.923 – 1.210	14.572	**<0.001**
Semantic [unrelated]	-0.040	0.010	-0.059 – -0.021	-4.134	**<0.001**
Gender [incongruent]	-0.002	0.010	-0.021 – 0.016	-0.257	0.797
Random Effects					
σ^2	0.05				
$\tau_{00\ Item}$	0.00				
$\tau_{00\ Participant}$	0.01				
ICC	0.17				
$N_{\ Participant}$	18				
$N_{\ Item}$	26				
Observations	1469				
Marginal R^2 / Conditional R^2	0.006 / 0.178				

Table 4.C.2: *Model of the best fit for combining naming latencies in both English-Dutch and Chinese-Dutch groups, including estimated means, standard error, and confidence intervals (n = 38).*

Formula: naming latency ~ gender (congruent vs. incongruent) + semantic (related vs. unrelated) + group (Chinese-Dutch vs. English-Dutch) + (1| participant) + (1| item)					
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	1.123	0.069	0.988 – 1.258	16.273	<0.001
Gender [Incongruent]	0.004	0.008	-0.011 – 0.019	0.577	0.564
Semantic [Unrelated]	-0.042	0.008	-0.057 – -0.027	-5.474	**<0.001**
Group [English-DutchL2]	-0.049	0.084	-0.214 – 0.115	-0.590	0.555
Random Effects					
σ^2	0.06				
$\tau_{00\ Participant}$	0.01				
$\tau_{00\ Item}$	0.00				
ICC	0.15				
$N_{\ Participant}$	37				
$N_{\ Item}$	26				
Observations	3189				
Marginal R^2 / Conditional R^2	0.014 / 0.162				

4.D Permutation test outcome

Figure 4.D.1: *Permutation test outcome for the English-Dutch L2 group (n = 18). Larger F-values are shown in darker colors and indicate a likelihood for statistically semantic relatedness (above panel) and gender congruency (below panel) effects on voltage amplitudes.*



4.E Model parameters: EEG results in the English-Dutch group

Table 4.E.1: *The best-fit model for N400 voltage amplitudes for the English-Dutch group, including estimated means, standard error, confidence intervals, and t-values (n = 18).*

Formula: amplitude ~ gender (congruent vs. incongruent) + semantic (related vs. unrelated) + hemisphere (left vs. right vs. midline) + (gender * semantic | participant) + (1| item)

Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	2.566	1.127	0.356 – 4.775	2.276	0.023
Gender [Incongruent]	-0.420	0.459	-1.319 – 0.480	-0.915	0.360
Semantic [Unrelated]	-1.138	0.502	-2.122 – -0.153	-2.265	**0.024**
Hemisphere [midline]	-0.714	0.048	-0.808 – -0.620	-14.861	<0.001
Hemisphere [right]	-0.567	0.048	-0.661 – -0.473	-11.793	<0.001

Random Effects

σ^2	162.84
$\tau_{00\ Item}$	3.96
$\tau_{00\ Participant}$	20.28
$\tau_{11\ Participant.SemanticUnrelated}$	10.24
$\tau_{11\ Participant.GenderIncongruent}$	10.40
$\tau_{11\ Participant.SemanticUnrelated:GenderIncongruent}$	35.03
$\rho_{01\ Participant.SemanticUnrelated}$	-0.47
$\rho_{01\ Participant.GenderIncongruent}$	-0.29
$\rho_{01\ Participant.SemanticUnrelated:GenderIncongruent}$	0.10
ICC	0.15

$N_{Participant}$	18
N_{Item}	26
Observations	422988
Marginal R^2 / Conditional R^2	0.002 / 0.148

Table 4.E.2: *The best-fit model for N400 voltage amplitudes for high-proficiency English-Dutch speakers (n = 3), including estimated means, standard error, confidence intervals, and t-values.*

Formula: amplitude ~ gender (congruent vs. incongruent) * semantic (related vs. unrelated) + hemisphere (left vs. right vs. midline) + (semantic | participant) + (1| item)

Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	2.599	1.871	-1.068 – 6.267	1.389	0.165
Gender [Incongruent]	-2.400	0.094	-2.584 – -2.215	-25.492	**<0.001**
Semantic [Unrelated]	-2.571	0.661	-3.865 – -1.276	-3.892	**0.029**
Hemisphere [midline]	-0.764	0.080	-0.922 – -0.607	-9.528	<0.001
Hemisphere [right]	0.903	0.080	0.745 – 1.060	11.252	<0.001
Gender [Incongruent] × Semantic [Unrelated]	2.251	0.132	1.993 – 2.509	17.108	**<0.001**

Random Effects

σ^2	89.70
$\tau_{00\ Item}$	10.57
$\tau_{00\ Participant}$	9.27

$\tau_{11\ Participant.SemanticUnrelated}$	1.28
$\rho_{01\ Participant}$	-0.33
ICC	0.18
$N_{Participant}$	3
N_{Item}	26
Observations	83628
Marginal R^2 / Conditional R^2	0.015 / 0.190

4.F Model parameters: EEG results in the combination analysis

Table 4.F.1: *The best-fit model for N400 voltage amplitudes in the combination analysis, including estimated means, standard error, confidence intervals, and t-values (n = 37).*

Formula: amplitude ~ gender (congruent vs. incongruent) * semantic (related vs. unrelated) * group (English-Dutch vs. Chinese-Dutch) + hemisphere (left vs. right vs. midline) + (gender * semantic | participant) + (1| item)

Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	1.430	0.952	-0.436 – 3.297	1.502	0.140
Gender [Incongruent]	-0.720	0.467	-1.636 – 0.196	-1.540	0.132
Semantic [Unrelated]	-0.867	0.052	-0.969 – -0.765	-16.635	**<0.001**
group [English-Dutch]	0.047	1.302	-2.504 – 2.598	0.036	0.971
Hemisphere [midline]	-1.877	0.033	-1.941 – -1.812	-57.312	<0.001
Hemisphere [right]	-0.434	0.033	-0.498 – -0.369	-13.244	<0.001
Gender [Incongruent] ×Semantic [Unrelated]	0.538	0.073	0.394 – 0.682	7.339	**<0.001**
Gender [Incongruent] ×group [English-Dutch]	0.149	0.670	-1.165 – 1.463	0.222	0.825

Semantic [Unrelated] ×group [English-Dutch]	-0.446	0.077	-0.596 – -0.296	-5.830	**<0.001**
(Gender [Incongruent] ×Semantic [Unrelated]) ×group [English-Dutch]	-0.098	0.107	-0.309 – 0.113	-0.910	0.363

Random Effects

σ^2	162.36
$\tau_{00\ Participant}$	15.63
$\tau_{00\ Item}$	2.14
$\tau_{11\ Participant.GenderIncongruent}$	4.10
$\rho_{01\ Participant}$	-0.26
ICC	0.10
$N_{\ Participant}$	37
$N_{\ Item}$	26
Observations	908697
Marginal R^2 / Conditional R^2	0.005 / 0.103

Table 4.F.2: *The best-fit model for P600 voltage amplitudes in the combination analysis, including estimated means, standard error, confidence intervals, and t-values (n = 37).*

Formula: amplitude ~ gender (congruent vs. incongruent) * group (English-Dutch vs. Chinese-Dutch) + semantic (related vs. unrelated) * group (English-Dutch vs. Chinese-Dutch) + hemisphere (left vs. right vs. midline) + (gender * semantic | participant) + (1| item)

Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	3.816	1.329	1.211 – 6.421	2.871	0.004
Gender [Incongruent]	0.161	0.659	-1.131 – 1.453	0.244	0.809
group [EnDuL2]	2.844	1.835	-0.752 – 6.440	1.550	0.130
Semantic [Unrelated]	0.929	0.588	-0.223 – 2.082	1.580	0.123
Hemisphere [midline]	0.347	0.033	0.282 – 0.411	10.551	<0.001
Hemisphere [right]	-1.006	0.033	-1.070 – -0.941	-30.582	<0.001
Gender [Incongruent] ×group [EnDuL2]	-0.760	0.945	-2.612 – 1.093	-0.804	0.427
group [EnDuL2] × Semantic [Unrelated]	-1.332	0.843	-2.985 – 0.320	-1.580	0.123

Random Effects

σ^2	244.78
$\tau_{00\ Participant}$	33.79
$\tau_{00\ Item}$	3.34

τ_{11} Participant.SemanticUnrelated	18.78
τ_{11} Participant.GenderIncongruent	14.87
τ_{11} Participant.SemanticUnrelated:GenderIncongruent	33.13
ρ_{01} Participant.SemanticUnrelated	-0.44
ρ_{01} Participant.GenderIncongruent	-0.53
ρ_{01} Participant.SemanticUnrelated:GenderIncongruent	0.28
ICC	0.13
$N_{Participant}$	37
N_{Item}	26
Observations	1358547
Marginal R^2 / Conditional R^2	0.005 / 0.134