Original Article

second language research

Second Language Research
1–48

DOI: 10.1177/02676583241227709
journals.sagepub.com/home/slr
Sage

Cross-linguistic effects in grammatical gender assignment and predictive processing in L1 Greek, L1 Russian, and L1 Turkish speakers of Norwegian as a second language

Janne Bondi Johannessen
University of Oslo, Norway

Björn Lundquist
Yulia Rodina
UiT The Arctic University of Norway, Norway

Eirik Tengesdal
University of Oslo, Norway
OsloMet – Oslo Metropolitan University, Norway

Nina Hagen Kaldhol
University of California San Diego, USA

Emel Türker
University of Oslo, Norway

Valantis Fyndanis
Cyprus University of Technology, Cyprus
University of Oslo, Norway

Corresponding author:
Björn Lundquist, UiT Norges arktiske universitet, Hansine Hansens veg 18, Tromsø, Troms 9019, Norway.
Email: bjorn.lundquist@uit.no

Abstract

The present study examines grammatical gender knowledge in offline production (gender marking on indefinite articles) and online gender processing (visual world paradigm) in adult second language (L2) learners of Norwegian with three different first languages (L1s): Greek, Russian, and Turkish. In particular, it investigates the role of the following factors: (1) presence vs. absence of grammatical gender in L1 (Norwegian, Greek and Russian have gender, whereas Turkish does not), (2) lexical gender congruency, (3) structural similarity between L1 and L2 in the realization of gender, and (4) proficiency in L2. In offline production, no difference was found between the three L2 groups: they all overused the default gender (masculine). However, L1 effects were observed in the eye-tracking task, where the high-proficiency L1 Greek and L1 Russian speakers showed earlier and more prominent signs of predictive gender processing compared to the high-proficiency L1 Turkish speakers. There were no effects of lexical gender congruency or structural similarity. This suggests that, when it comes to predictive gender processing, what matters is proficiency and the presence vs. absence of grammatical gender in the L1. We interpret the findings in the context of current approaches to predictive processing emphasizing the role of cue reliability and utility.

Keywords

cross-linguistic effects, grammatical gender assignment, Greek, lexical gender congruency, Norwegian, predictive processing, Russian, structural/syntactic congruency, Turkish

I Introduction

Grammatical gender is a linguistic phenomenon where a great degree of cross-linguistic variation can be observed. Not all languages express grammatical gender, and the gender systems that exist differ in terms of number of gender values and their labels, variety and transparency of gender assignment cues expressed by nouns (if any), and variety and transparency of structural cues. Grammatical gender has also been shown to be hard to acquire for adult second language (L2) learners. Recent research on grammatical gender in late L2 acquisition has directed specific attention to how lexical similarities in gender assignment and syntactic similarities in the realization of gender agreement between first language (L1) and second language (L2) affect L2 gender learning and processing (see, amongst others, Alemán Bañón et al., 2014; Dussias et al., 2013; Grüter et al., 2012; Hopp and Lemmerth, 2018; Loerts, 2012; Sabourin and Stowe, 2008). The question is no longer whether gender features may be universally available to L2-learners regardless of whether their L1 expresses gender, but rather how the activation of gender proceeds based on shared lexical and/or syntactic correspondences between the L1 and the L2. In other words, it is of special interest how L2 learners deploy their grammatical knowledge and how they detect and access the morphosyntactic mechanisms in their two co-existing languages in real time. This line of research proposes to focus on cross-linguistic effects and the role of the language-internal factors that characterize the gender systems in the L1 and the L2. The core question is thus if L1–L2 similarities in their gender systems facilitate L2 gender acquisition and processing and, in such case, what kind of similarities are likely to facilitate or interfere with the L2 grammar and processing, e.g.

similarities in syntactic patterns, lexical groupings (i.e. number of gender categories) or morphological realization.

The findings so far are inconclusive and come from a limited set of language pairs, e.g. L1 Romance (French, Italian, Spanish) and L1 German / L2 Dutch (Sabourin and Stowe, 2008), L1 Polish / L2 Dutch (Loerts, 2012), L1 Italian and L1 English / L2 Spanish (Dussias et al., 2013) and L1 Russian / L2 German (Hopp and Lemmerth, 2018). There is some evidence that L2 learners activate the grammatical gender of the L1 in L2 production (picture naming) and comprehension (Bordag and Pechmann, 2007; Klassen, 2016; Paolieri et al., 2010; Weber and Paris, 2004). The effects of structural similarity have mostly been studied independently and less systematically. Some studies demonstrate that structural overlap between the L1 and L2 may advantage L2 learners to the extent that they can exhibit nativelike grammatical gender processing in the L2 not only at advanced, but also at intermediate proficiency levels (Hopp and Lemmerth, 2018). Mixed results are reported for L2 gender production (oral and written), where some studies observe facilitative effects for speakers of gendered languages, while others do not (Sabourin et al., 2006 vs. Ragnhildstveit, 2017).

The present study aims to contribute to a more comprehensive understanding of how grammatical gender is acquired during adult L2 learning by examining how gender is assigned in oral production and how gender knowledge is implemented in online gender processing. Our main goal is to take a granular approach to the effects of lexical and structural similarities and differences between the L1 and L2 gender systems in second language acquisition (SLA). We extend the scope of research to the previously unstudied language pairs L1 Greek / L2 Norwegian, L1 Russian / L2 Norwegian and L1 Turkish / L2 Norwegian, which exhibit a varying degree of overlap in gender properties. Although Norwegian, Greek and Russian categorize nouns into one of the three gender classes (masculine, feminine or neuter), they differ in the lexical grouping of the gender classes, i.e. not all individual nouns are assigned the same gender across these languages. At the syntactic level, Norwegian and Greek mark gender on indefinite articles, in contrast to Russian, which lacks indefinite articles. Turkish on the other hand lacks grammatical gender altogether. Thus, these language pairs provide a good testing ground for the effects of L1–L2 similarity in gender assignment and online gender processing in the L2.

1 Grammatical gender in late L2 acquisition: L1 effects in production and predictive processing

Grammatical gender has been the subject of extensive research in late L2 acquisition across different languages showing that even advanced learners can have protracted difficulties with gender. In spoken production, difficulties manifest themselves in the form of overgeneralization errors; for example, masculine is often erroneously overgeneralized with neuter nouns in L2 Norwegian (e.g. *en eple* 'an apple' instead of *et eple*; Anderssen and Busterud, 2022). In online processing, even advanced L2 learners appear to be unable to use gender marking as a facilitative cue to activate upcoming nouns (Grüter et al., 2012). Some previous theoretical models define L1 transfer effects of grammatical gender as all-or-nothing effects where the existence of grammatical gender

in the L1 determines the degree of target-likeness in the L2. According to the Full Transfer / Full Access / Full Parse Model, an abstract gender feature can transfer at the initial state of adult L2 acquisition and facilitate the acquisition and processing of gender (Dekydtspotter et al., 2006). Similarly, in the Competition Model, grammatical gender is readily available in the L2 for learners whose L1 has gender (MacWhinney, 2008). In the absence of gender in the L1, L2 learners may have difficulty representing a new abstract grammatical feature in a target L2 (Representational Deficit Hypothesis; Hawkins and Chan, 1997; Tsimpli and Dimitrakopoulou, 2007). The Feature Reassembly Hypothesis argues for a stepwise process of (re-)assembling abstract gender features in the course of L2 acquisition (Lardiere, 2009).

Recent research by Hopp and Lemmerth (2018) argues that the effects of L1 are more fine-grained, and the degree of difficulty with gender is affected interactively by proficiency and the amount of overlap in L1 and L2 gender realization, both at the lexical and syntactic level. This proposal is based on a detailed analysis of L1 Russian / L2 German learners' gender processing and the comparison of this group with a group of L1 English / L2 German learners reported in Hopp (2013, 2016). Hopp and Lemmerth (2018) observe that both lexical and syntactic congruency facilitate predictive gender processing but their interactive effects are only evident in the high-intermediate proficiency learners (advanced learners show nativelike predictive gender processing across all contexts). Specifically, in a syntactically congruent condition (agreement on adjectives), gender marking in German was used predictively by the high-intermediate proficiency group irrespective of whether the nouns were congruent in Russian and German. In contrast, in a syntactically incongruent condition (agreement on articles), predictive gender processing obtained only for congruent nouns. This suggests that less proficient L2 learners fail to inhibit the activation of L1 lexical representations when they compete with the lexical representations from the L2 in contexts in which there is a lack of structural overlap between L1 and L2.[1] Based on these results, Hopp and Lemmerth argue that, in high-intermediate-proficiency learners, problems with L2 predictive gender processing are due to the linguistic properties of the L1 and arise in contexts in which L1 does not afford predictive processing.

While Hopp and Lemmerth's (2018) proposal is novel and needs to be tested across other L1–L2 combinations, their findings are consistent with previous research showing that predictive gender processing is modulated by learner proficiency and L1 properties. L2 learners whose L1 encodes gender, e.g. L1 Italian / L2 Spanish speakers in Dussias et al. (2013), could exploit gender on articles to facilitate the processing of the following noun despite their low L2 proficiency. Dussias et al. (2013) proposed that the presence of gender in L1 may have a modulating effect on gender processing in L2, especially when the gender systems of the two languages exhibit lexical and morphosyntactic similarities. In contrast, L1 speakers of a language without grammatical gender (such as English) can show predictive effects of gender marking on determiners only at high-proficiency L2 levels. In Dussias et al. (2013), only high-proficiency L1 English / L2 Spanish speakers showed evidence of predictive gender processing. Nevertheless, not always do high-proficiency L2 learners exhibit predictive gender processing (see Grüter

et al., 2012). The failure to predict received special attention in the seminal paper by Kaan and Grüter (2021) who emphasized the role of cue reliability and utility in predictive processing. This approach originates in the Competition Model (Bates and MacWhinney, 1981; MacWhinney and Bates, 1989) according to which languages weigh cues differently: what is a reliable cue in the L1 may be regarded as a reliable cue in the L2 if the cues are shared between the two languages. When there is no straightforward overlap between the languages, some cues may not be reliable for an L2 speaker because their representations are not sufficiently specified or entrenched. It is this approach that will guide our understanding of the findings in the present study.

For gender assignment in production, even advanced L2 learners exhibit non-target-like behavior, especially if their L1 does not have gender (e.g. Franceschina, 2005). Although the Spanish gender system is highly transparent and offers learners reliable phonological cues for gender assignment on nouns, advanced L1 English / L2 Spanish learners only reach 75%–90% accuracy in elicited production (Alarcón, 2011; Bruhn de Garavito and White, 2003; Franceschina, 2005; Grüter et al., 2012; Montrul et al., 2008). In languages with opaque gender systems, such as German, gender assignment among late learners is shown to be even more variable with accuracy ranging from 53% to 100% (Hopp, 2013). Performance on gender assignment in L2 Dutch by speakers of German, Romance, and English led Sabourin et al. (2006) to the conclusion that having grammatical gender in the L1 is an important factor in L2 gender assignment, but the L2 gender system has to be similar to that of the L1 in order for the L2 gender distinctions to be mastered. However, the evidence from the acquisition of L2 Norwegian by speakers of Vietnamese, English, German, Spanish, and Dutch in Ragnhildstveit (2017) does not support this conclusion. In that study, the L1 groups' accuracy performance on gender marking on indefinite articles ranged between 82% and 88%, and there were no significant between-group differences.

A separate line of research has focused on lexical gender congruency effects, investigating whether gender congruent stimuli (i.e. target L2 nouns having the same gender as the translation equivalent nouns in the L1) can facilitate gender selection in the L2 due to the activation of the L1 gender nodes which are shared between the two languages. There is a large body of evidence obtained from bare noun and determiner phrase naming showing that naming response is facilitated when the L2 target noun and its L1 translation equivalent noun have the same gender, but is inhibited when the L2 and L1 equivalent nouns have different genders (Bordag, 2004; Bordag and Pechmann, 2007; Klassen, 2016; Lemhöfer et al., 2008; Morales et al., 2011; Paolieri et al., 2010). Lexical gender congruency effects have also been reported in written- and spoken-word recognition tasks during which L1 lexical representations could spread activation of their corresponding L2 lexical entries (Morales et al., 2016; Salamoura and Williams, 2007). Such results have been argued to support the Gender Integrated Representation Hypothesis (Salamoura and Williams, 2007), according to which gender nodes are shared between languages. In the present study, we aim to determine how congruence or incongruence in gender between Norwegian and Greek or Russian affects L2 gender assignment and online ('real-time') processing.

2 Grammatical gender in Norwegian, Greek, Russian, and Turkish

A Norwegian. Traditionally, Norwegian is regarded as having three genders: masculine, feminine and neuter (Faarlund et al., 1997). Masculine is considered to be the lexical default, as masculine nouns constitute the majority of nouns in Norwegian. According to Trosterud (2001), masculine nouns make up 52%, feminine nouns 32%, and neuters only 16% (based on frequency counts in the Nynorsk Dictionary (Hovdenak, 1998)). Gender assignment is traditionally viewed as non-transparent, as the nouns themselves do not provide reliable gender cues (Rodina and Westergaard, 2017). Within the noun phrase (NP), gender is marked on indefinite articles, attributive adjectives, possessive pronouns, demonstratives and prenominal determiners in the so-called double definite forms (for a detailed description of the gender system of Norwegian, see Rodina and Westergaard, 2015). The present study investigates gender production and processing of indefinite articles only. Importantly, while various dialects of Norwegian still have a traditional three-gender system, several dialects are in the process of losing the feminine, including the dialects where data collection took place, Oslo and Tromsø (Fretheim, 1985 [1976]; Lødrup, 2011; Lundquist and Vangsnes, 2018; Rodina and Westergaard, 2021). Therefore, in the present study, we focus on two stable genders: masculine and neuter, illustrated in (1). In the experiments, we only use nouns that have been traditionally masculine and neuter in the dialects spoken in Oslo and Tromsø.

(1)	a.	en	bil	(Oslo and Tromsø dialects)
		a.M	car(M)	
	b.	et	hus	
		a.N	house(N)	

B Greek. Greek nouns are masculine, feminine or neuter. In contrast to Norwegian and Russian, neuter is the most frequent gender in Greek, whereas masculine is the least frequent one (e.g. Salamoura and Williams, 2007). Gender marking is expressed prenominally on articles and adjectives as well as other targets. Greek gender is considered to be phonologically transparent, since the noun endings *-as*, *-a*, and *-o* are associated with masculine, feminine and neuter, respectively (Anastassiadis-Symeonidis and Chila-Markopoulou, 2003; Hulk, 2017).

(2)	a.	énas	pínakas	(Greek)
		a.M	painting(M)	
	b.	mɲa	vraðjá	
		a.F	evening(F)	
	c.	éna	vivlío	
		a.N	book(N)	

C Russian. Russian has a three-gender system of masculine, feminine and neuter. Based on dictionary counts, the frequency of masculine nouns in Russian is 46%, while there is 41% feminine and 13% neuter nouns (Corbett, 1991). Gender assignment is rather transparent, as noun endings provide unambiguous cues in most cases. Gender is expressed

only in the singular on adjectives, possessive and demonstrative pronouns, as well as verbs in the past tense. Russian does not have definite or indefinite articles.

(3)	a.	bol′šoj	dom	(Russian)
		big.M	house(M)	
	b.	bol′šaja	mašina	
		big.F	car(F)	
	c.	bol′šoje	okno	
		big.N	window(N)	

D Turkish. Turkish is a morphologically rich language, but it does not have gender or an article system. Note however that the numeral *bir* 'one' (*iyi bir ev* 'a/one good house'), can be used as an indefinite article (Göksel and Kerslake, 2005; Lewis, 2000 [1967]).

In sum, the target gender system of L2 Norwegian investigated in the present study has lexical and syntactic similarities and differences compared to the gender systems of L1 Greek and L1 Russian. At the lexical level, all three languages distinguish masculine and neuter as well as feminine, which however was not included in the study due to its unstable status in the Oslo and Tromsø dialects. Across these languages some nouns are assigned the same gender in the L1 and the L2, while others are not. At the syntactic level, Greek is more similar to Norwegian than Russian, since it marks gender on articles, in contrast to Russian, which lacks articles.

II The present study

The present study investigates the proposal that the effects of L1 in L2 grammatical gender assignment and predictive processing are fine-grained and that the degree of the overlap between the gender systems in the L1 and L2, as well as L2 proficiency and L1–L2 lexical gender congruency, determine the extent to which gender assignment in offline production and online gender processing in the L2 is nativelike (Dussias et al., 2013; Hopp and Lemmerth, 2018). To investigate this proposal, we focus on L2 Norwegian speakers from three L1 backgrounds: Greek, Russian, and Turkish. The present study addresses three main research questions:

- Research question 1: Do L1–L2 lexical gender congruency and morphosyntactic similarity facilitate grammatical gender assignment and online processing in the L2?
- Research question 2: Does the absence of gender in the L1 affect L2 learners' gender assignment and online processing in the L2? If so, how does it affect them?
- Research question 3: To what extent does L2 proficiency modulate grammatical gender assignment and online processing in the L2?

We expect the gender system of Norwegian to be challenging for all our L2 learners, since it is highly non-transparent and unstable. The overview of the morphosyntactic similarities and differences presented in Section I.2 suggests that, although both Greek and Russian distinguish masculine and neuter noun classes, also found in Norwegian, the

overlap with L2 Norwegian is highest for Greek, as it marks gender on articles, while Russian does not. We predict, therefore, that gender-marked articles will facilitate gender assignment and predictive processing in L1 Greek / L2 Norwegian learners, who may outperform L1 Russian / L2 Norwegian learners. At the same time, the L1 Turkish / L2 Norwegian learners should have no advantage, since Turkish does not express grammatical gender. Furthermore, we predict that gender assignment and processing will be modulated by proficiency. In light of the evidence in Hopp and Lemmerth (2018) and Dussias et al. (2013), we can expect that nativelike gender processing may obtain for L1 Greek learners already at lower proficiency levels due to the structural similarities between Greek and Norwegian. However, gender prediction may be problematic even for high proficiency learners of genderless languages, like Turkish. Finally, on the assumption that gender nodes are shared across languages, as suggested by the Gender Integrated Representation Hypothesis (Salamoura and Williams, 2007), we predict that L1 Greek and L1 Russian learners will activate the lexical gender of the L1 during L2 gender assignment/production and online processing and demonstrate more target-like performance in lexically congruent (e.g. Russian: *jabloko*(N) 'apple'; Norwegian: *eple*(N) 'apple') than in lexically incongruent trials (e.g. Russian: *dom*(M) 'house'; Norwegian: *hus*(N) 'house').

III Method

The study includes four tasks which were performed in the same sequence by all participant groups. Experiment 1 was an object naming task, which elicited noun forms preceded by indefinite articles in Norwegian. This task was followed by an eye-tracking experiment (visual world paradigm), Experiment 2, which consisted of two sessions with a short break in between during which the participants completed a background survey. Experiment 2 was followed by a Norwegian language proficiency test. Finally, the participants performed the same object naming task as in Experiment 1 in their L1s. This was necessary to ensure that the participants could activate the intended target nouns in their L1s. The complete experimental session took approximately 60 minutes per participant.

1 Participants

The participants were 66 L2 speakers of Norwegian, divided into three groups: 23 L1 Greek, 23 L1 Russian, and 20 L1 Turkish. They were recruited in Oslo and Tromsø. Most of them had higher education (BA, MA, PhD). Some had upper-secondary, post-secondary or technical education. Their occupations varied across all groups (teachers, janitors, researchers, doctors, university students, etc.); none were linguists. We had two control groups of L1 Norwegian speakers. Control Group 1 ($n=19$, age range=25–55 years) was tested on the materials created for L1 Greek and L1 Turkish speakers. Control Group 2 ($n=14$, age range=20–25 years) was tested on the materials created for L1 Russian speakers.

Table 1 summarizes the background information of the L2 learners, including their self-assessment of Norwegian competence on a scale from 1 to 10 (speaking and listening) and their self-assessment of how often they spoke and read in Norwegian. Table 1

Table 1. Background information of the second language (L2) participants.

	L1 Greek	L1 Russian	L1 Turkish
Number of participants	23 (18 female)	23 (18 female)	20 (14 female)
Mean age (range) (years)	40 (27–64)	43 (28–64)	49 (32–65)
Mean AoA (range) (years)	30 (18–50)	30 (19–50)	25 (18–31)
Mean LoR (range) (years)	10 (1–46)	13 (2–24)	24 (2–38)
L2 listening, self-assessment (range)	7.8 (5–10)	7.2 (3–10)	8.3 (4–10)
L2 speaking, self-assessment (range)	7.2 (4–10)	7 (2–10)	7.8 (4–10)
Speak the L2 on a daily basis (*n*/total)	22/23	20/23	20/20
Read in the L2 on a daily basis, (*n*/total)	21/23	20/23	20/20
Proficiency test, proportion correct (range) (Proficiency Measure 1)	0.97 (0.75–1.00)	0.92 (0.69–1.00)	0.94 (0.8–1.00)
Object naming, proportion correct (range) (Proficiency Measure 2)	0.83 (0.69–0.95)	0.83 (0.61–1.00)	0.92 (0.58–1.00)
Neuter score, proportion correct (range) (Proficiency Measure 3)	0.62 (0.25–0.90)	0.66 (0.28–0.90)	0.69 (0.28–0.97)
Composite Proficiency Measure 1	0.81 (0.55–1.00)	0.77 (0.48–1.00)	0.87 (0.47–1.00)
Composite Proficiency Measure 2	0.51 (0.13–0.84)	0.51 (0.16–0.85)	0.61 (0.18–0.94)

Notes. AoA = Age of Arrival (years). LoR = Length of Residence (years), self-assessment on a scale from 1 to 10.

shows that nearly all L2 learners spoke Norwegian on a daily basis and assessed themselves as proficient speakers of Norwegian. There is one clear between-group difference: the mean Length of Residence (LoR) for L1 Turkish speakers is 24 years, whereas the mean LoR of L1 Greek and L1 Russian speakers is 10 and 13 years, respectively.

The background data also contain three proficiency measures that were considered in the study. Proficiency Measure 1 is a multiple-choice proficiency test (a placement test for L2 learners of Norwegian designed at UiT The Arctic University of Norway), which included 36 questions covering morphological and syntactic knowledge, among other domains (see Appendix 1). All groups performed at ceiling (92%–97%) in this test. Proficiency Measure 2 is the proportion of correctly named nouns in Experiment 1. Proficiency Measure 3 is gender assignment accuracy with neuter nouns in Experiment 1 (neuter score). In all three measures, the proficiency distribution is highly left-skewed (Proficiency Measure 1: −1.5 skewness; Proficiency Measure 2: −0.67 skewness; Proficiency Measure 3: −0.55 skewness). Since none of the above measures is able to capture the more nuanced between-participant variation in proficiency, we used composite proficiency scores in Experiment 1 and Experiment 2. Since the neuter score is one of the dependent variables in Experiment 1, the composite proficiency score used in Experiment 1 (Composite Proficiency Measure 1) is based on Proficiency Measure 1 and Proficiency Measure 2 only. The composite proficiency score used in Experiment 2 was based on all three measures. Both composite proficiency scores were calculated by multiplying their component proficiency measures. This procedure gives a proficiency scale from 0 to 1 (Composite Proficiency Measure 1: mean=0.81, skewness=−0.82, range=0.47–1; Composite Proficiency Measure 2: mean=0.54, skewness=−0.13, range=0.14–0.94).

2 Materials and procedure

The stimuli in Experiments 1 and 2 consisted of 64 images of objects depicting common inanimate nouns (see Appendix 2). All images were designed specifically for this study to ensure that the style and color range were similar across all of them. The selected nouns were all unambiguously picturable. Cognates and nouns with many salient and frequent synonyms were excluded. We avoided using nouns that were feminine in any dialect of Norwegian. The nouns were classified based on gender and L1–L2 gender congruency: congruent neuter (16), incongruent neuter (16), congruent masculine (16), and incongruent masculine (16). Since it was impossible to match the nouns for gender and congruency across Norwegian, Greek and Russian, two sets of nouns were created. One set was used with the L1 Greek and L1 Turkish learners, as well as with Control Group 1. The second set was used with the L1 Russian learners and Control Group 2. Twenty out of 64 nouns were replaced in this set and the lexical gender congruency values were different for the Greek and Russian groups in 21 of the overlapping nouns. The frequency of the nouns in the Greek/Turkish set ranged from 188 to 186,356 lemma tokens (mean: 18,693) and in the Russian set from 188 to 242,680 (mean: 21,955) in the approximately 700-million word corpus of written Bokmål Norwegian (Norwegian Web as Corpus; Guevara, 2010).

In Experiment 1, the participants, including the Norwegian control participants, were asked to name objects shown on a screen by saying, e.g. *Jeg ser en ballong* ‘I see a balloon’. Errors with indefinite articles were not corrected, but if a participant failed to

Table 2. Experiment 2: The eye-tracking design.

Target	Competitor	Congruency	Target gender	Competitor gender
			L2 (L1)	L2 (L1)
Masculine (32)	Same (16)	Congruent (8)	M (m)	M (m)
		Incongruent (8)	M (n)	M (n)
	Different (16)	Congruent (8)	M (m)	N (m)
		Incongruent (8)	M (n)	N (n)
Neuter (32)	Same (16)	Congruent (8)	N (n)	N (n)
		Incongruent (8)	N (m)	N (m)
	Different (16)	Congruent (8)	N (n)	M (n)
		Incongruent (8)	N (m)	M (m)

Notes. In the Target gender and Competitor gender columns, the L2 Norwegian gender values are given in capital letters and the gender values of the L1 translation equivalents are given in parentheses. M = masculine. N = neuter.

name the depicted object, the experimenter provided the noun, preceded by the correct article. This was done to make sure that the participants knew all test nouns before the eye-tracking experiment. The test items were randomized for each participant by SMI Experiment Center. The experiment was not timed. The answers were audio recorded and later transcribed.

In Experiment 2, the participants saw two objects on the screen: target and competitor, which were either of the same gender in Norwegian (Same Condition) or of different gender (Different Condition) (Table 2). Each test noun appeared twice as a target, once in the Same and the other time in the Different Condition, and twice as a competitor. In total, Experiment 2 consisted of 128 experimental trials and 128 fillers, i.e. 256 displays in total. The target appeared the same number of times on the right- and left-hand side. Two lists were made where the location of the target was altered for each item. The eye-tracking was split into two parts, which were of the same length and contained the same number of items in each condition. The filler panels also consisted of two images from the same pool, but contained unrelated auditory cues (see below).

The lexical gender congruency set-up in Table 2 and Figures 1 and 2 shows that the L2 target nouns and their L1 translation equivalents were always gender-matched in the congruent condition, and gender-mismatched in the incongruent condition. The L1 translation equivalent of the competitor had the same gender as the L1 translation equivalent of the target.[2] This design makes it possible to investigate the effect of lexical gender congruency in online processing: if congruency matters, more looks to the target image are expected in congruent than incongruent trials, i.e. in trials where the target noun has the same gender in the L1 and L2. Moreover, the rationale behind this set-up was to ensure that predictive looks to the target are triggered by the knowledge of the L2 gender. Since the target and the competitor always had the same gender in the corresponding L1 nouns, the participants could not rely on L1 gender knowledge to locate the target noun.

The eye-tracking study (Experiment 2) immediately followed the production study (Experiment 1). The participants had thus been familiarized with the images prior to the eye-tracking study. The auditory stimuli consisted of the carrier phrase *Jeg tenker på* . . .

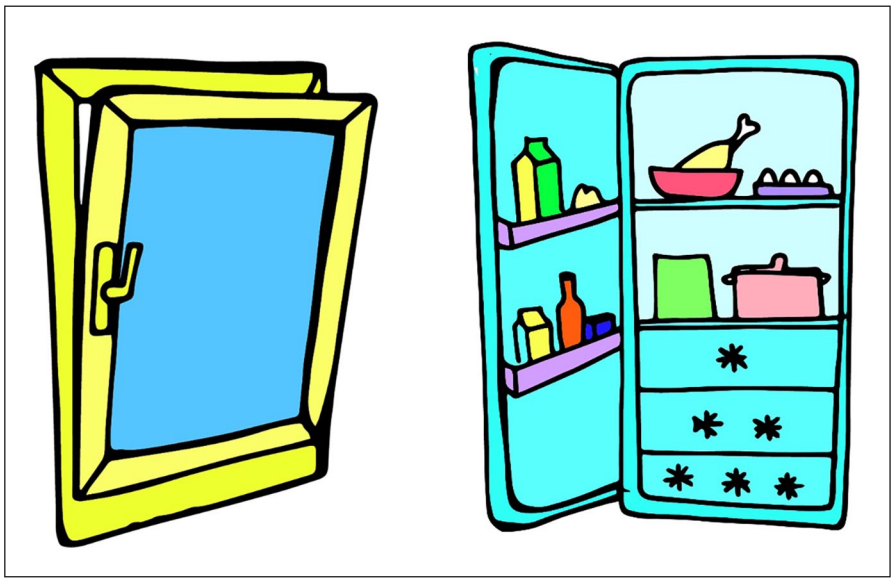


Figure 1. Same congruent panel for the Norwegian–Greek language pair. Target (right): *fridge*: Norwegian *kjøleskap*(N); Greek *psijío*(N). Competitor (left): *window*: Norwegian *vindu*(N); Greek *paráθiro*(N).

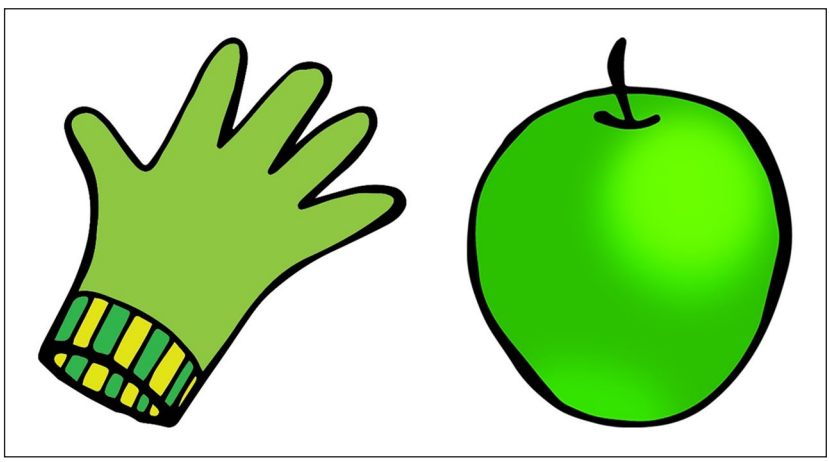


Figure 2. Different incongruent panel for the Norwegian–Greek language pair. Target (left): *glove*: Norwegian *hanske*(M); Greek *γá(n)di*(N). Competitor (right): *apple*: Norwegian *eple*(N); Greek *mílo*(N).

'I am thinking of . . .' followed by an NP consisting of a gender-inflected indefinite article, followed by an adjective with no gender marking and the target noun (*en/et avbilda* NOUN 'a(M/N) depicted NOUN'). The uninflected adjective had the function of increasing the time between article offset and noun onset and to ensure that the article and noun

Table 3. Experiment 2: Timing of the experimental item (ms indicate onsets).

Carrier phrase	Article	Adjective	Target noun
0 ms	920 ms	1,192 ms	1,980 ms
Jeg tenker på	*en/et*	*avbilda*	*bil/tog* (1,980 →)
'I am thinking about'	'a'(M)/'a'(N)	'depicted'	'car'(M)/'train'(N)

were not processed by the participants as one unit (see Brouwer et al., 2017; Grüter et al., 2012). A female native speaker of the Oslo dialect provided the auditory stimuli recorded in a sound-proof room using a Zoom Handy Recorder H4n. Items representing a normal speech rate and naturalistic prosody were selected for the study. Using Praat (Boersma, 2001), the recordings were segmented and transcribed with TextGrids, and the carrier phrase was acoustically adapted to ensure that it was identical for all utterances to avoid any effects of unintended segmental and suprasegmental cues. Half of the items included the masculine indefinite article *en* and the other half the neuter *et*. In the carrier phrases, the onset of articles and adjectives was consistently at 920 ms and 1,192 ms, respectively. The onset of the target noun was consistently at 1,980 ms after the onset of the carrier phrase (Table 3). Mean noun duration of the target noun was 552 ms (range: 336–1,008 ms). All audio files were leveled to the Heavy degree and with a −70 dB noise threshold in Audacity® to increase homogeneity (Audacity Team, 2015). In filler carrier phrases, the target nouns were substituted by plural nouns, e.g. *Jeg tenker på to avbilda objekter* 'I am thinking of two depicted objects'. The numeral *to* 'two' was inserted in place of the article, ensuring that the onsets were identical.

Eye-tracking data from 53 participants (23 Greek, 17 Turkish, and 13 Norwegian speakers in Control Group 1) were collected in Oslo with an SMI RED250 mobile eye-tracker with a sampling rate of 250 Hz. Data from 46 participants (23 Russian, three Turkish, six Norwegian speakers in Control Group 1, and 14 Norwegian speakers in Control Group 2) were collected in Tromsø with an SMI RED500 eye-tracker at a sampling rate of 250 Hz. Participants sat on a chair with a distance of 50–60 cm. The tracking equipment was calibrated by instructing the participant to fixate on five points on the screen and validated twice during the experiment, once before each eye-tracking list. The ideal calibration angle was $\leqslant 0.5°$. For a minority of participants, a lower accuracy was accepted after two failed attempts of recalibration. There were only two areas of interest, each covering approximately 45% of the stimulus screen. Although the screens for the two eye-trackers were not of the same size, the ratio of the size of the images to the size of the screen was the same. To begin each trial, participants looked at a fixation point in the center of the computer screen. Subsequently, participants saw two pictures and listened to a carrier phrase which was played simultaneously. A new trial appeared automatically; there was no clicking or manual selection of the target item involved.

IV Results

1 Norwegian object naming task: Experiment 1

The results of Experiment 1 are presented in Table 4. All three L2 Norwegian groups correctly named the majority of the test objects (Measure 1) and scored high on gender

Table 4. Results of experiment 1 (mean accuracy in percentage with range in parentheses).

	L1 Greek	L1 Turkish	L1 Russian
Measure 1: Correctly named objects	83 (68–93)	92 (58–100)	83 (60–100)
Measure 2: Gender assignment accuracy	74 (49–93)	77 (53–94)	74 (43–93)
Measure 3: Gender assignment accuracy by gender value	M: 87 (50–100) N: 63 (25–90)	M: 85 (34–97) N: 70 (26–97)	M: 82 (41–100) N: 66 (28–90)
Measure 4: Gender assignment accuracy by lexical gender congruency	C: 81 (50–97) I: 69 (35–93)	C: 81 (56–100) I: 73 (48–94)	C: 73 (39–93) I: 76 (48–94)

Notes. Measure 1: percentage of objects named correctly. Measure 2: gender assignment accuracy with correctly named objects. Measure 3: gender assignment accuracy with masculine (M) and neuter (N) nouns. Measure 4: gender assignment accuracy with lexically congruent (C) and incongruent (I) nouns.

assignment with the correctly named objects (Measure 2). Neuter gender was more problematic than masculine for all groups (Measure 3). Gender assignment with congruent nouns seemed more target-like than gender assignment with incongruent nouns for L1 Greek, but not for L1 Russian. The L1 Turkish participants, who were tested on the same nouns as the L1 Greek, also performed better with the Greek–Norwegian congruent nouns than with the Greek–Norwegian incongruent ones.

We explored the results by fitting a series of logistic mixed-effects regression models using the R packages *afex* (Singmann et al., 2016) and *lme4* (Bates et al., 2015), with Gender Assignment Accuracy as the dependent variable. Since the L1 Russian group was tested on a partially different set of nouns than the L1 Greek and L1 Turkish groups (see Section III.2), we fitted separate models to the two data sets (Greek/Turkish and Russian).

The model fitted to the Greek/Turkish data set included four main predictors:

- L1 (two levels: Greek, Turkish);
- Norwegian–Greek Lexical Gender Congruency (two levels: congruent, incongruent);
- L2 Proficiency (continuous variable centered at mean); and
- Norwegian Target Gender (two levels: masculine, neuter).

Although Turkish lacks gender, the Turkish data were also coded for congruency, with the corresponding Norwegian–Greek congruency values. In addition to the interaction between L1 and Lexical Gender Congruency, we included in the model two three-way interactions targeting the relationship between (a) L2 Proficiency, Norwegian–Greek Lexical Gender Congruency and L1, and (b) L2 Proficiency, Norwegian Target Gender and L1, as well as their nested two-way interactions. The model included random intercepts for Participant and Item. The model is shown in (4) and the full regression table is provided in Appendix 3.

(4) Production Model: Gender Accuracy ~ (L1 × L2 Proficiency × Lexical Gender Congruency) + (L1 × L2 Proficiency × Norwegian Target Gender) + (Norwegian Target Gender × Lexical Gender Congruency) + (1 | Participant) + (1 | Item), family = binomial)

Given the nested structure of the data and because we are interested in higher level interactions (up to three levels), we will not report the values of the coefficients and their standard errors, but only the chi-square statistics obtained from likelihood ratio tests carried out in the *afex* package in R (Singmann et al., 2016). Coefficients and standard errors are reported in the text when the interpretation of the chi-square statistics is not obvious in light of the raw results.

For the Greek and Turkish L1-groups, we found:

- a significant main effect of Norwegian Target Gender ($\chi^2(1) = 27.10$, $p < .001$): there were more errors with neuter than with masculine nouns, indicating defaulting to masculine;
- a significant main effect of Lexical Gender Congruency ($\chi^2(1) = 10.21$, $p < .01$): there were overall more errors with the incongruent nouns than with the congruent nouns; and
- a significant main effect of L2 Proficiency ($\chi^2(1) = 13.11$, $p < .001$): the higher the participants' L2 Proficiency, the higher their gender assignment scores.

There was, furthermore, a significant interaction between L2 Proficiency and Norwegian Target Gender ($\chi^2(1) = 24.07$, $p < .001$): the effect of L2 Proficiency was smaller for masculine than for neuter nouns, indicating that the lower proficiency participants used masculine as a default gender value to a greater extent compared to the higher proficiency speakers.

Importantly, there was no significant main effect of L1 and no two- or three-ways interactions involving L1. Crucially, there was no significant interaction between L1 and Lexical Gender Congruency ($\chi^2(1) = 1.84$; $p = .174$) and no three-way interaction between L1, Lexical Gender Congruency, and L2 Proficiency ($\chi^2(1) = 0.04$; $p = .84$). This suggests that the main effect of Lexical Gender Congruency is an artifact of the material rather than a true congruency effect. If there were a true gender congruency effect, this effect should be restricted to the Greek group, and be fully absent in the L1 Turkish group, which here can be seen as a control group for the lexical gender congruency variable.

A closer look at the results reveals that five incongruent neuter nouns caused problems for both the L1 Greek and L1 Turkish participants (*kamera* 'camera', *anker* 'anchor', *basseng* 'pool', *jordbær* 'strawberry', and *fengsel* 'prison'), as these nouns elicited below 50% accuracy performance for both groups. Figure 3 visualizes the effects of Lexical Gender Congruency and Norwegian Target Gender for individual nouns in the L1 Greek and L1 Turkish groups. Each dot represents an item in the test, color and shape-coded for congruency and Norwegian gender, and the regression line shows correlation between the Greek and Turkish responses ($r^2 = 0.52$, $p < .001$). The graph shows that the two L1 groups struggle with mainly the same nouns, and that gender congruency per se is not an obvious mitigating factor for the Greek L1 group (i.e. there is no interaction between L1 and Congruency).

The regression model fitted to the L1 Russian data set included *Norwegian Target Gender* (masculine, neuter), *Norwegian–Russian Lexical Gender Congruency*

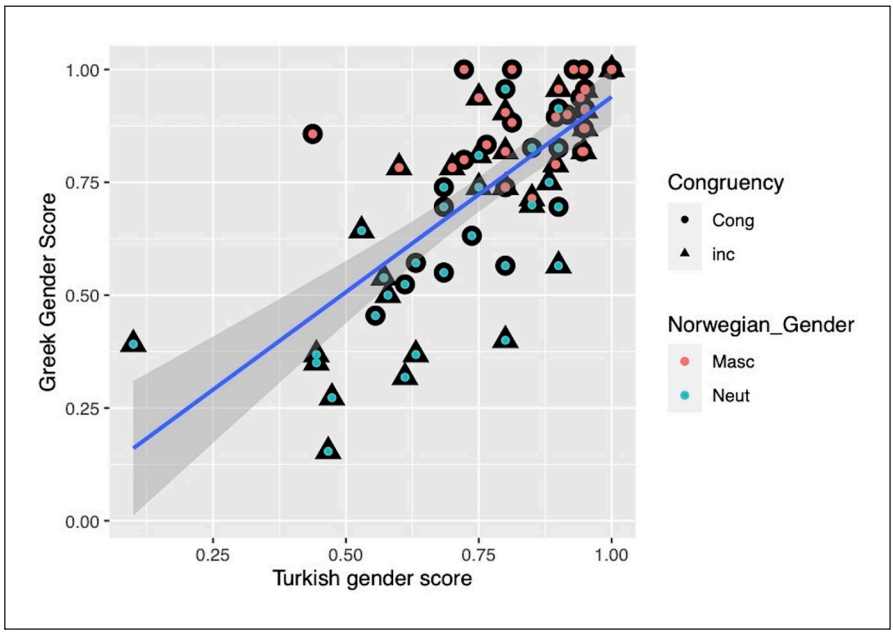


Figure 3. Effects of lexical gender congruency and Norwegian target gender (masculine, neuter) on gender assignment with individual nouns in the first language (L1) Greek and L1 Turkish groups.
Note. Regression line shows correlation between the Greek and Turkish gender results.

(congruent, incongruent), *L2 Proficiency* (continuous variable) and all interactions between these three predictors as fixed effects, as well as random intercepts for Participants and Items. Similarly to the results of the L1 Greek and L1 Turkish groups, there was a significant main effect of L2 Proficiency ($\chi^2(1)=9.04, p<.01$) and Norwegian Target Gender ($\chi^2(1)=16.07$, $p<.001$), with neuter nouns eliciting significantly more errors than masculine nouns. There was no significant main effect of Norwegian–Russian Lexical Gender Congruency ($\chi^2(1)=0.69$, $p=.406$). The three-way interaction between Norwegian Target Gender, Norwegian–Russian Lexical Gender Congruency and L2 Proficiency was not significant ($\chi^2(1)=0.08$, $p=.775$).[3] The results from this model are presented in Appendix 4.

In sum, the results of the production task reveal a significant effect of Norwegian Target Gender across all L2 groups with neuter being more problematic than masculine and a significant effect of L2 Proficiency on gender assignment in all three groups as well as the interaction between L2 Proficiency with Norwegian Target Gender. The L1 Greek and L1 Turkish groups both performed worse with incongruent than congruent nouns and, given the absence of gender in Turkish, we interpret this finding as an unforeseen effect of the choice of nouns rather than an effect of congruency per se. No gender congruency effect was found in the L1 Russian group. These results suggest that L2 learners tend to resort to masculine rather than to the gender of the corresponding noun in their L1 in production.

2 Online gender processing: Experiment 2 (eye-tracking)

In the eye-tracking study, we are interested in how the Same/Different gender manipulation affects looks to target in the three L1 groups across the whole proficiency range. In addition, we investigate if L1–L2 gender congruent target nouns attract more looks than the incongruent nouns. We modelled the proportion of looks to target in two 50 ms-long temporal regions of interests (RoIs), an early RoI and a late RoI, which were determined based on the analysis of the eye-tracking data from Control Group 1 (Greek/Turkish data set) and Control Group 2 (Russian data set). The first RoI (Early RoI) is the earliest temporal region in which the control groups showed a reliable effect of the Same/Different manipulation (Condition). The second RoI (Late RoI) was defined as the last temporal region in which the fixation patterns of the control groups were guided by the gender on the article, and not by the onset of the lexical noun. This was diagnosed by the divergence of the proportion of looks to target and competitor in the Same gender condition (see where the red and grey lines diverge in Figure 4): as fixations cannot be guided by gender information in the same condition, a target-over-competitor advantage reveals that fixations are guided by the lexical noun. As illustrated in Figure 4, the Early RoI is right before the noun onset for Control Group 1 (Proportion of looks (PoL) to target, Same: 0.25, PoL to target, Different: 0.33, $p < .01$) and 100 ms prior to the noun onset for Control Group 2 (PoL to target, Same: 0.44, PoL to target, Different: 0.6, $p < .01$). The late RoI is 450 ms after the noun onset for Control Group 1 and 350–400 ms after the noun onset for Control Group 2.[4] For an analysis of the proportion of looks across the whole trial time, and not just for the two regions of interest, see also Appendix 9.

In the models, the dependent variable was ±fixation at target coded as 1 (fixation at target) and 0 (fixation at a competitor or white space). The models fitted were mixed-effects logistic regressions (*lme4*; Bates et al., 2015) with random intercepts for Item and Participant and a by-Participant slope for the effect of Condition. We were interested in the effect of five predictors and their interactions: Condition (Same/Different gender), L1 (Greek, Russian, Turkish), Norwegian Target Gender (masculine, neuter), L1–L2 Lexical Gender Congruency (congruent, incongruent), and L2 Proficiency. The latter was treated as a continuous variable based on three proficiency measures (Composite Proficiency Measure 2; see Section III.1) and was centered at the mean. To avoid problems with the interpretation and convergence of models with higher-level interactions, we fitted separate models for the main and interaction effects of Condition, given in (5), and the main and interaction effects of Congruency, given in (6). The model for Condition includes all predictors and their interactions except for congruency. To facilitate the presentation and interpretation of the results, we also fitted individual models for each of the L1s.[5] In addition, we will plot the proportion of looks to target for the individual L1 and proficiency groups for easy comparison between the test groups and control groups (Figure 4). The model for congruency excluded the main and interaction effects of Norwegian gender. To test any subtle effects of congruency, we fitted individual models for two L1s with grammatical gender (Greek and Russian), with L1–L2 Lexical Gender Congruency, L2 Proficiency, Condition, and the interactions between these as predictors.

(5) Condition (same/different) model: FixationTarget ~ Condition × L1 × L2 Proficiency × Norwegian Target Gender + (1 + Condition|Participant) + (1|Item)

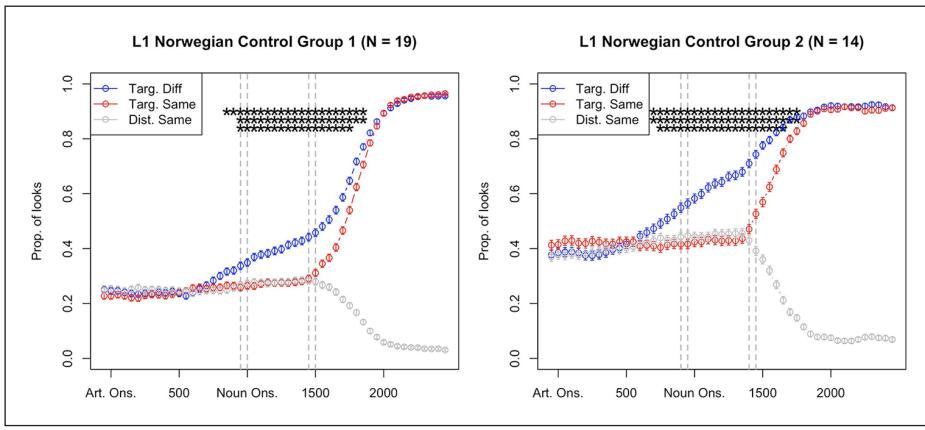


Figure 4. Eye-tracking results of Control Group 1 (left panel, Greek/Turkish data set, $n = 19$) and Control Group 2 (right panel, Russian data set, $n = 14$).
Notes. The blue line represents the proportion of looks to target in the Different condition. The red line shows the proportion of looks to the target in the Same condition. The grey line shows the proportion of looks to the competitor (distractor) in the Same condition. The early and late RoI (50 ms time slots) are marked with dotted lines. The asterisks above the lines represent significance values based on a series of logistic mixed-effects models (lme4, emmeans).

(6) L1–L2 Lexical Gender Congruency (congruent/incongruent): FixationTarget ~ L1 × L1–L2 Lexical Gender Congruency × L2 Proficiency × Condition + (1 + Condition|Participant) + (1|Item)

All eye-tracking trials were included in the analysis except for the trials with tracking loss within the two RoIs. In total 7,610 observations in the Early RoI and 7,902 in the late RoI were included in the model. We did not exclude trials based on the production results, since we do not presuppose a direct link between production and comprehension. Indeed, a by-item analysis revealed no strong correlations between production scores and predictive looks at either the Early or Late RoI (both r^2 values below 0.1). For additional motivation of the choice of trial inclusion, see Appendix 9.

Below, we present chi-square statistics for the main and interaction effects obtained from model comparisons using the *afex* package (Singmann et al., 2016) and *lme4* (Bates et al., 2015) in R (see regression tables in Appendices 5 and 6). Beta-coefficients from the full model or L1-specific smaller models are given when the interpretation of the chi-square statistics is not straightforward. Table 5 shows the chi-square statistics for Condition, L1, L2 Proficiency, Norwegian Target Gender, and their interactions. At the early RoI, there was a significant interaction between Condition and L2 Proficiency, and a significant three-way interaction between Condition, L2 Proficiency, and L1. There were no significant interactions involving Norwegian Target Gender and no significant main effect of this variable. There was a marginally significant effect of Condition and a significant main effect of L1: the groups differed in how much they fixated on the target irrespective of the Same/Different Condition. The post-hoc tests for the individual L1s

Table 5. Effects of condition, first language (L1), second language (L2) proficiency and Norwegian target gender at early and late regions of interest (RoIs).

		Early RoI		Late RoI	
Effect	df	χ^2	*p*	χ^2	*p*
Condition	1	3.03	.082	14.85	<.001***
L2 proficiency	1	0.00	.981	2.45	.118
L1	2	6.85	.032*	4.84	.089
Norwegian target gender	1	0.00	.971	2.69	.101
Condition: L2 proficiency	1	11.34	<.001***	27.53	<.001***
Condition: L1	2	0.46	.793	1.84	.399
L2 Proficiency: L1	2	3.99	.136	7.34	.026*
Condition: Norwegian target gender	1	0.37	.542	0.07	.794
L2 Proficiency: Norwegian target gender	1	0.00	.955	0.03	.858
L1: Norwegian target gender	2	1.38	.500	1.46	.482
Condition: L2 proficiency: L1	2	6.40	.041*	6.38	.041*
Condition: L2 proficiency: Norwegian target gender	1	0.63	.428	2.69	.101
Condition: L1: Norwegian target gender	2	1.42	.492	3.56	.168
L2 Proficiency: L1: Norwegian target gender	2	2.10	.351	0.67	.715
Condition: L2 proficiency: L1: Norwegian target gender	2	0.05	.975	0.54	.765

Notes. *$p < 0.05$. ***$p < 0.001$.

are presented in Tables 6 to 8. We found a significant interaction between Condition and L2 Proficiency for both L1 Greek and L1 Russian (Tables 6 and 8, respectively), but not for L1 Turkish (Table 7).

At the late RoI, there was a significant main effect of Condition, a significant interaction between Condition and L2 Proficiency, and a significant three-way interaction between Condition, L2 Proficiency, and L1 (Table 5). Post-hoc tests (individual glmers for the three L1 groups) revealed significant interactions between Condition and L2 Proficiency for L1 Greek (Table 6) and L1 Russian (Table 8), but not for L1 Turkish (Table 7). These models also revealed a significant main effect of Condition in the L1 Greek group (Table 6), but not for the other two groups. However, the omnibus model (Table 5) showed no significant interaction between Condition and L1, and the beta-coefficients for Condition were positive for both the L1 Russian (Table 8) and L1 Turkish groups (Table 7), indicating more looks to target in the different gender condition for all groups. Finally, the significant interaction between L2 Proficiency and L1 (Table 5) was driven by a negative effect of L2 Proficiency in the L1 Turkish group (see Table 7).[6]

3 Time-course analysis and visualizations

To clearly illustrate the differences between the three L1 groups at different L2 proficiency levels and to facilitate comparisons between them and the native speakers, we

Table 6. Eye-tracking results of the first language (L1) Greek group. Condition is dummy coded (intercept is Same) and L2 Proficiency is a continuous variable centered at mean.

	Early RoI (2,601 observations, $n = 23$)			Late RoI (2,714 observations, $n = 23$)		
	β	*SE*	*p*	β	*SE*	*p*
Intercept	−0.77	0.19	< .001***	−0.65	0.19	< .001***
Condition	0.16	0.11	.14	0.37	0.1	< .001***
L2 proficiency	−0.39	1.06	.71	0.56	1.06	.59
Condition: L2 proficiency	1.34	0.66	.042*	2.37	0.61	< .001***

Notes. *$p < .05$. ***$p < .001$.

Table 7. Eye-tracking results of the first language (L1) Turkish group.

	Early RoI (2,368 observations, $n = 20$)			Late RoI (2,463 observations, $n = 20$)		
	β	*SE*	*p*	β	*SE*	*p*
Intercept	−0.44	0.13	< .001***	−0.28	0.11	< .01**
Condition	0.06	0.09	.49	0.17	0.1	0.12
L2 proficiency	−1.1	0.52	.039*	−1.22	0.45	< .01**
Condition: L2 proficiency	0.11	0.36	.76	0.76	0.48	.11

Note. Condition is dummy coded (intercept is Same) and L2 Proficiency is a continuous variable, centered at mean. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 8. Eye-tracking results of the first language (L1) Russian group.

	Early RoI (2,641 observations, $n = 23$)			Late RoI (2,725 observations, $n = 23$)		
	β	*SE*	*p*	β	*SE*	*p*
Intercept	−0.20	0.11	.06	−0.12	0.12	.3
Condition	0.11	0.1	.27	0.19	0.1	.052
L2 proficiency	−0.06	0.6	.9	0.12	0.64	.85
Condition: L2 proficiency	1.59	0.54	< .001***	2.27	0.56	< .001***

Note. Condition is dummy coded (intercept is Same) and L2 Proficiency is a continuous variable, centered at mean. ***$p < .001$.

also provide time course graphs for the effect of the Same/Different Condition (Figures 5–7). In these graphs, proficiency is not treated as a continuous variable; instead, the three L1 groups were divided into high-proficiency and intermediate-proficiency subgroups using the median split (i.e. above or below the median L2 proficiency score for each L1 group). Significance markers in the graphs are based on logistic mixed-effects

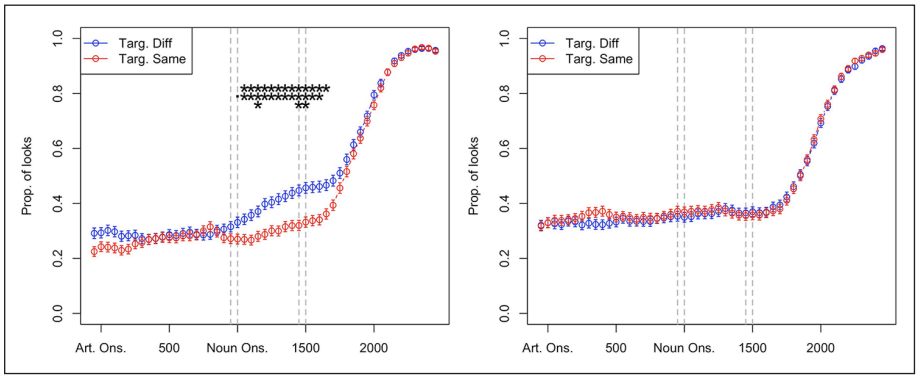


Figure 5. Gender prediction in first language (L1) Greek high-proficiency subgroup, $n = 11$ (left) and intermediate-proficiency subgroup, $n = 12$ (right).
Note. Asterisks are to be read vertically: *$p < .05$. **$p < .01$. ***$p < .001$.

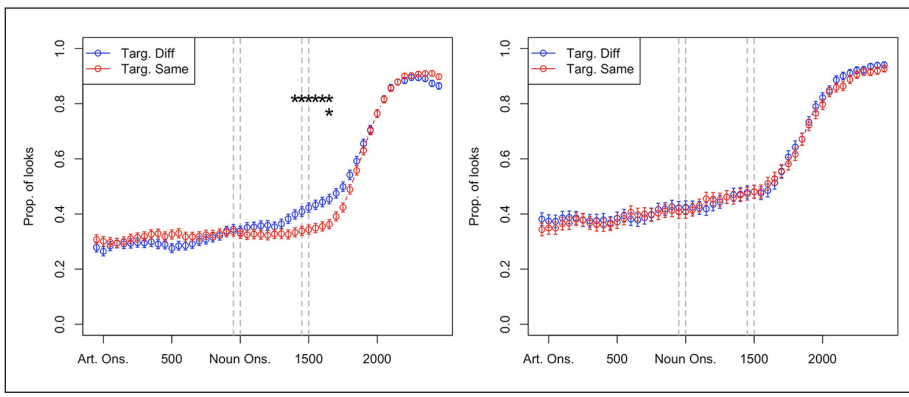


Figure 6. Gender prediction in first language (L1) Turkish high-proficiency subgroup, $n = 10$ (left) and intermediate proficiency subgroup, $n = 10$ (right).
Note. Asterisks are to be read vertically: *$p < .05$. **$p < .01$.

regression models with pairwise comparisons between the Same and Different Condition values for each subgroup (categorical, median-split coding of proficiency), retrieved from the *emmeans* package (Lenth, 2021). For alternative illustrations with proficiency as a continuous measure, see Appendix 9.

In all intermediate-proficiency L1 groups, proportions of looks to target in the Same and Different condition were more or less identical throughout the trials, suggesting that looks to the target were not guided by the gender-marked article at any point in time. In the high-proficiency L1 groups, interactions between L1 and Condition emerged. Both the L1 Greek (Figure 5) and L1 Russian (Figure 7) high-proficiency groups showed an effect of Condition at the early RoI, and this effect was maintained until the late RoI. In both groups, the temporal signature was similar to that of the L1 control groups: the

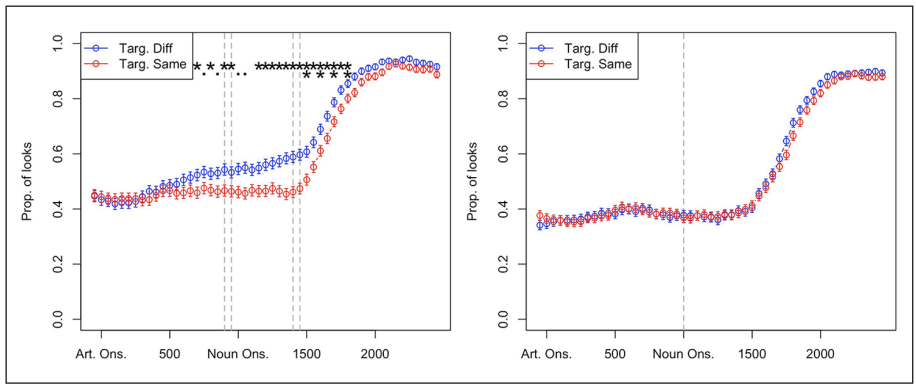


Figure 7. Gender prediction in first language (L1) Russian high-proficiency subgroup, $n = 12$ (left) and intermediate proficiency subgroup, $n = 11$ (right).
Note. Asterisks and points are to be read vertically: *$p < .05$. **$p < .01$. ***$p < .001$.

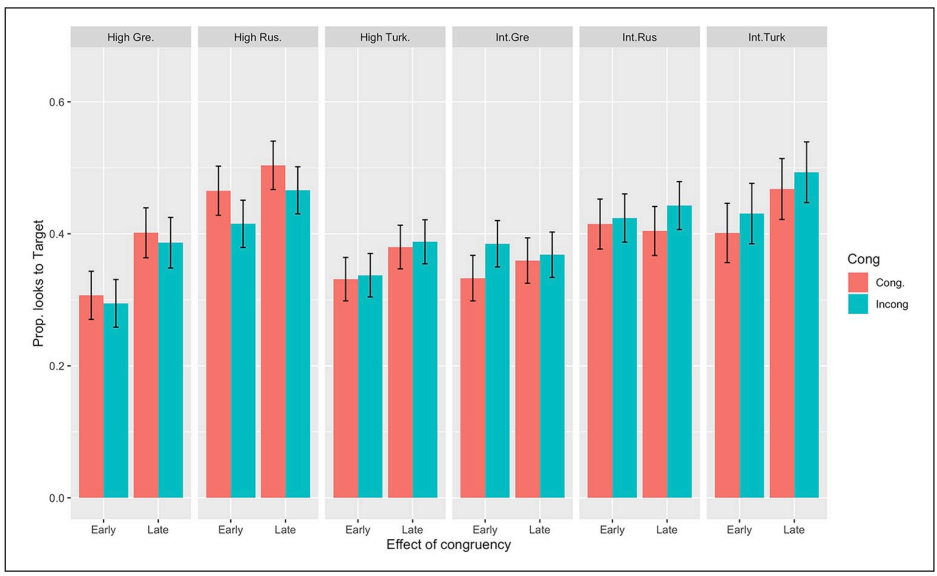


Figure 8. Lexical gender congruency effects in high- and intermediate-proficiency learners for the early and late region of interest (RoI).

separation of the two lines (Same/Different condition) took place at roughly the same point in time for the test and control groups. However, there was no effect of Condition at the early RoI in the L1 Turkish high-proficiency group; the separation of the two lines does not take place until just before the late RoI.

4 L1–L2 lexical gender congruency

In the omnibus model that targeted congruency, we found no main or interaction effects related to congruency in either the early or late RoI. To make sure that no subtle interactions were lost in the big model, we fitted individual models for the three L1s. The chi-square and *p*-values are presented in Appendix 7. Here, we found an interaction between Lexical Gender Congruency and L2 Proficiency in the L1 Greek group at the early RoI ($\chi^2(2) = 5.27$, $p = .022$). This interaction is hard to interpret in the absence of main effects of Lexical Gender Congruency. Nevertheless, the group-level analyses suggest that intermediate-proficiency L1 Greek participants are more likely to have early fixations at incongruent targets ($\beta = 0.3$, $SE = 0.18$, $p = .087$), while high-proficiency L1 Greek participants show no difference between the congruent and incongruent targets. The effects of Lexical Gender Congruency for high- and intermediate-proficiency subgroups at the early and late RoI are illustrated in Figure 8.[7] In contrast to Experiment 1 where the L1 Greek and L1 Turkish participants performed better on Norwegian–Greek gender-congruent than on gender-incongruent nouns, in the eye-tracking task, the gender-congruent nouns did not attract more early looks than the gender-incongruent nouns, suggesting that the relationship between the production and processing results on an item level is by no means straightforward (as was already evidenced by the lack of correlation between production and eye-tracking results, reported in Section IV.1).

V Discussion

The research questions formulated in the present study asked whether and how the production and processing of grammatical gender in the L2 are affected by L2 proficiency and three linguistic variables:

- degree of structural similarity of the gender systems of L1 and L2;
- ±presence of a gender system in the L1; and
- ±overlap in lexical gender classification in L1 and L2.

The focus of investigation was the masculine–neuter distinction in L2 Norwegian, a language with a non-transparent gender system. To our knowledge, this is the first study which compares the effects of L1–L2 lexical and structural similarities and differences in L2 learners with three different L1s: Greek, Russian, and Turkish. Greek has close structural similarities with the target L2 Norwegian in that both languages mark masculine and neuter prenominally on indefinite articles. Russian has gender but lacks articles and Turkish has no gender.

We first analyzed the effects of L1–L2 similarities and differences in an object naming task which tapped into L2 learners' knowledge of gender in Norwegian. Despite our expectations, we found no differences between the three learner groups. For all groups, we found an effect of L2 proficiency: the most proficient speakers showed evidence of near-target-like knowledge of grammatical gender in Norwegian, while speakers of

lower proficiency experienced difficulties. Like monolinguals and L2 learners of Norwegian from other studies (Anderssen and Busterud, 2022; Rodina and Westergaard, 2015), our participants experienced more problems with the neuter, and their errors revealed overgeneralization of masculine which decreased with increased proficiency in Norwegian.

In contrast to previously observed facilitative effects for speakers of gendered languages (Sabourin et al., 2006), we found no effect of L1 in the production study. There was no evidence of structural facilitation (predicted for L1 Greek participants followed by L1 Russian participants) or delay (predicted for L1 Turkish participants) in the production data. Furthermore, the facilitating effect of lexical gender congruency observed in previous research (e.g. Bordag and Pechmann, 2007; Paolieri et al., 2010) is not obvious in the present data. Even though gender assignment in the L1 Greek participants was facilitated by gender congruent stimuli, the performance of L1 Turkish participants was also better with exactly the same nouns, and the L1 Russian participants, who were tested on a partially different noun set, were not facilitated by L1–L2 lexical gender congruency. It thus appears that the main effect of lexical gender congruency observed in the L1 Greek and L1 Turkish data sets was an artifact of the experimental material. It is still unclear what caused the difference between the Norwegian–Greek congruent and incongruent nouns for the L1 Greek and L1 Turkish participants. Future research on lexical gender congruency should consider including L2 speakers with a genderless L1 as a control group in order to rule out confounding variables in the test items (see, for example, Morales et al., 2016).

From our results, it is clear that the masculine–neuter distinction can be acquired by L1 speakers of languages with or without grammatical gender, and the absence of evidence of facilitation or delay from L1 suggests that gender assignment in L2 Norwegian is not determined by the L1/L2 similarities and differences but rather by the characteristics of the target gender system. Interestingly, our findings are different from those of Sabourin et al. (2006) who report a gradient performance in L2 Dutch, with L1 German outperforming L1 Romance and L1 English learners. While Norwegian and Dutch are similar in that they both have non-transparent gender systems distinguishing masculine and neuter, it is hard to draw a direct comparison here, since the gender systems in the participants' L1s in the two studies are not the same.

Having examined the gender knowledge of our L2 learners, we then investigated whether they would use this knowledge in real-time speech processing. In the visual world eye-tracking task, the three L1 groups performed partly differently. For participants of a lower L2 proficiency and a limited knowledge of the L2 gender system, we found no signs of predictive gender processing in any of the L1 groups. As for the participants of higher L2 proficiency, L1 Greek and L1 Russian participants performed similarly to the L1 Norwegian control groups and were able to use the gender cues on the indefinite articles to predict the upcoming nouns. The high-proficiency L1 Turkish participants, on the other hand, differed in their performance from the high-proficiency L1 Greek and L1 Russian participants, as well as from the L1 Norwegian controls: the effect of the Same/Different manipulation was smaller and, most importantly, it was only seen very late in the eye-tracking trial, about 400 ms after the noun onset, in contrast to the high-proficiency L1

Greek and L1 Russian participants who showed evidence of predictive gender processing already prior to the noun onset. These results suggest that the properties of the L1 grammatical gender systems may have repercussions on L2 gender but, crucially, this is only seen in perception and not in production. Our results suggest that L2 gender processing with an L1-like temporal profile is more likely to be found in L2 speakers with grammatical gender in their L1 compared to L2 speakers with a genderless L1. Still, predictive gender processing can be found even in L2 learners whose L1 does not express gender, although the effects may be smaller or delayed compared to learners with gender in their L1.

In our view, these differential outcomes are not compatible with all-or-nothing accounts and the representational deficit position in particular, which assumes that L1 and L2 acquisition and processing are fundamentally different (Hawkins and Chan, 1997; Tsimpli and Dimitrakopoulou, 2007). Instead, we would like to argue that abstract linguistic knowledge of L2 gender is present in all our learners, including L1 Turkish speakers, but it is not implemented in the same way during online processing by speakers of different L1s. Following Kaan and Grüter (2021), we propose that the answer to the question why learners with considerable L2 experience and relevant linguistic knowledge fail to use this knowledge in real-time speech processing is in the differential reliability of the relevant cues across language systems. The cue reliability and utility approach predicts that L2 speakers may place different weights on different cues and they may consider some cues as unreliable. When there is no straightforward overlap between the L1 and the L2, some cues may not be reliable for an L2 speaker because their representations are not sufficiently specified or entrenched. The L1 Norwegian control data in our study suggest that gender-marked indefinite articles are reliable cues for gender prediction in Norwegian. Unlike Norwegian, Turkish has no gender, and morphosyntactic marking in this language is largely postnominal and bound (suffixal). Given this, it may not be surprising that L1 Turkish learners of Norwegian do not find indefinite articles reliable enough to make predictions about upcoming nouns. After all, they do not have prior L1 experience with a grammatical feature that controls agreement on associated words (in our case, words preceding nouns). In other words, even though, when tested offline, L1 Turkish speakers can use *en* and *et* with respective nouns, they do not rely heavily on the gender information that these articles provide during real-time processing. This outcome may be due to the processing speed needed while listening, the uncertainty regarding what they hear, or due to less specified lexical representations (see the discussion in Kaan and Grüter, 2021).

To conclude, the abstract representations of grammatical gender are not deficient in L1 Turkish participants, as they performed equally well (or better) in the gender assignment task as the Greek and Russian speakers. Rather, gender representations are more entrenched in L1 Greek and L1 Russian participants who are used to establishing overtly marked agreement relations between a noun and its dependents in their L1s. Furthermore, the lack of differences between the L1 Greek and L1 Russian groups suggests that L1–L2 structural similarity is not the key issue, i.e. the presence of articles in the L1 is not a prerequisite for making predictions about upcoming nouns based on the gender values on articles in the L2. Based on the current language sample, we propose that it is sufficient

if the L1 overtly marks dependencies between the NP internal modifiers and lexical features inherent to the noun (i.e. gender).

VI Limitations of the study

In contrast to the core findings discussed above, other results from the study are less clear and harder to compare to previous research. First, given the L1–L2 lexical gender congruency set-up used in Experiment 2 (eye-tracking study), our results cannot be directly compared to those of Hopp and Lemmerth (2018). Since we did not include a condition where both target and competitor were gender-congruent in both the same and different condition, we cannot rule out the possibility that even intermediate-proficiency participants engage in predictive gender processing in their L2 when they have access to L1–L2 congruent gender information. Yet, as we did not find any lexical gender congruency effects in the production study and no signs of congruent nouns attracting more looks than incongruent nouns in the eye-tracking study, we find it unlikely that lexical gender congruency could facilitate predictive processing in intermediate-proficiency L2 speakers, but we cannot rule this out. We can however safely conclude that high-proficiency L1 Greek and L1 Russian speakers engage in predictive gender processing irrespective of L1–L2 gender congruency, in accordance with findings in Hopp and Lemmerth (2018).

Another limitation of our study is related to a potentially controversial choice of not excluding the individual eye-tracking trials in which participants failed to assign gender correctly. As pointed out by one of the reviewers, the fact that we found no predictive gender processing among the intermediate-proficiency participants may simply be due to the fact that they did not know the gender of the nouns. Our choice to include all trials was partly based on our skepticism towards equating the production results with the perceptive grammar; even advanced L2 learners can make occasional mistakes (slips) in their production. Moreover, as shown in Section IV.1, production scores were not correlated strongly with predictive looks at either the early or late RoI. Crucially, exclusions would lead to the disposing of a lot of eye-tracking data. As every noun/object appears in four trials (twice as a target and twice as a competitor), little data would remain for many of the participants (e.g. for someone scoring 75% correct in the production task, only 50% of the eye-tracking trials could be used). This would reduce the statistical power of the study and, more importantly, make comparisons between groups (test/control) and proficiency levels highly unreliable as the participants/groups would be tested on partly different sets of nouns.

Finally, we would like to point out that the implementation of the Visual World Paradigm in our study is quite robust against effects of mis-assignments of individual nouns. A mis-assignment of a noun would not lead to a reversed effect of the Same/Different gender manipulation, i.e. it would not lead to a Same-condition advantage, but only to a nullification of the manipulation. This means that we expect to find at least a small Different-over-Same advantage for all participants that have over 50% correct gender assignments in the production task, and no effect of the manipulation for the very few participants who scored below 50%, while a reversed effect never is expected. The graphs for the intermediate proficiency participants in Figures 5 to 7 clearly show that there is no indication of an effect of Condition at any point in time despite the fact that most of these participants scored well above 50% in the production task. This suggests that the intermediate proficiency participants do not make use of the gender cues for

predicting the upcoming noun, irrespective of their gender knowledge of the target and distractor nouns. For further discussion of this issue, see also Appendix 9.

Author note

The full dataset is openly available at: https://doi.org/10.18710/HFIF8K

Acknowledgements

This article is in honor of our deceased colleague Janne Bondi Johannessen who led and inspired the project. We would like to thank the people who discussed this project with us and helped us in various ways over several years: Yvonne Wilhelmina Henriëtte Van Baal, Ingrid Lossius Falkum, Franziska Köder, Marcel Nascimento de Moura, Linn Iren Sjånes Rødvand, Yeşim Sevinç, Tor A. Åfarli, Ruth Kramer, Terje Lohndal, Natalia Mitrofanova, Irina Sekerina, Marit Westergaard, Edith Kaan, and Jorge Valdés Kroff. Special thanks to the three anonymous reviewers and the editors of *Second Language Research*. We are very grateful to artist Tanja Russita for her wonderful drawings and cooperation. The authors gratefully acknowledge MultiLing's Socio-Cognitive Laboratory (UiO) and AcqVA Aurora Lab (UiT The Arctic University of Norway).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been funded in part by the Centre for Multilingualism in Society across the Lifespan (MultiLing) / the Research Council of Norway, project number 223265, and by the research project MultiGender at the Centre for Advanced Study at the Norwegian Academy of Science and Letters in Oslo during the academic year 2019–20.

ORCID iDs

Björn Lundquist https://orcid.org/0000-0002-1746-6769
Eirik Tengesdal https://orcid.org/0000-0003-0599-8925
Nina Hagen Kaldhol https://orcid.org/0009-0004-2208-0803
Valantis Fyndanis https://orcid.org/0000-0001-9403-3468

Notes

1. Intermediate-proficiency L2 English learners of German in Hopp (2013, 2016) were found not to use gender on German articles to predict upcoming referents. Since English and German both have articles (while Russian does not), Hopp and Lemmerth (2018) conclude that the difficulties in using L2 German articles for gender prediction by L1 Russian speakers cannot solely be due to the lack of articles in the L1.
2. There were some exceptions in the L1 Russian data set. As suitable Russian neuter nouns were hard to find, some of the competitors were feminine in Russian. As far as we can tell, this did not have an impact on the results.
3. We also fitted a model with all three L1s with the same structure as the model used for the Greek–Turkish data set. However, the model did not converge with a random intercept for Item, presumably due to the fact that the Russian data set was partly different from the

Greek–Turkish data set. The outcome of the model without Item as a random intercept was similar to the outcome of the Greek–Turkish model. The only significant effect of L1 emerged in a three-way interaction between L1, L2 Proficiency, and Norwegian Target Gender ($\chi^2(2) = 18.7$, $p < .001$): the Norwegian Target Gender by L2 Proficiency effect was smaller for L1 Russian than for L1 Greek and L1 Turkish groups. This may be an artifact of the material, as the five difficult neuter nouns were not part of the Russian data set.

4. There were overall more looks to white space (i.e. neither target nor competitor) in Control Group 1 than in Control Group 2 possibly because the control groups were tested on partially different test items and eye-trackers with screens of different sizes. Yet, it is not clear why the effect of the same/different manipulation is both larger and earlier in the Russian control group than the Greek–Turkish control group. Although it may appear counterintuitive to relate the RoIs to different time slots for the Greek–Turkish and the Russian groups, we believe that this is the correct thing to do, given the differences between the control groups. In Appendix 8, we explain the rationale behind including two RoIs, instead of applying, for example, a growth curve analysis (Mirman et al., 2008) or a divergence point analysis (Stone et al., 2021).
5. One of the reviewers pointed out a possible inconsistency in our article: in the production study we analysed the Russian results separately, as they were tested on a partly different set of items, but in the eye-tracking study we analysed the three L1 groups together. The reason for doing this is that the effect of item is very large in the production study, as was shown in Figure 3, in contrast to the eye-tracking study, where item differences account for only a very small portion of the variance. For full transparency, we provide an alternative analysis of the data in Appendix 9, where we analyse the Greek and Turkish results separately from the Russian results.
6. It is beyond the scope of this article to explore this effect. For some reason, the high-proficiency L1 Turkish participants looked more at white space (the fixation cross) than the intermediate-proficiency L1 Turkish participants, and none of them showed an effect of Condition.
7. There was a marginally significant three-way (Congruency × Condition × Proficiency) interaction for the L1 Russian group ($\chi^2 = 2.90$, $p = .089$) and the L1 Greek group ($\chi^2 = 2.89$, $p = .089$) at the early RoI. Since these results are only marginally significant and the p-values were not corrected for multiple comparisons, they give no direct evidence for a general effect of congruency.

References

Alarcón IV (2011) Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Bilingualism: Language and Cognition* 14: 332–50.

Alemán Bañón J, Fiorentino R, and Gabriele A (2014) Morphosyntactic processing in advanced second language (L2) learners: An event-related potential investigation of the effects of L1–L2 similarity and structural distance. *Second Language Research* 30: 275–306.

Anastassiadis-Symeonidis A and Chila-Markopoulou D (2003) Συγχρονικές και διαχρονικές τάσεις στο γένος της ελληνικής: Μια θεωρητική πρόταση [Synchronic and diachronic trends in the gender of Modern Greek: A theoretical proposal]. In: Anastassiadis-Symeonidis A, Ralli A, and Chila-Markopoulou D (eds) *Το γένος* [Gender]. Athens, Greece: Patakis, pp. 13–56.

Anderssen M and Busterud G (2022) Grammatisk kjønn og bøyningsklasse i norsk som andrespråk: En korpusstudie [Grammatical gender and declension in Norwegian as a second language: A corpus-based study]. *Norsk Lingvistisk Tidsskrift* 40: 87–127.

Audacity Team (2015) *Audacity®. Version 2.1.0*. [Audio editor and recorder]. Available at: http://audacityteam.org (accessed January 2024).

Bates D, Mächler M, Bolker B, and Walker S (2015) Fitting linear mixed-effects models using lm4. *Journal of Statistical Software* 6: 1–48.

Bates E and MacWhinney B (1981) Second language acquisition from a functionalist perspective: Pragmatic, semantic and perceptual strategies. In: Winitz H (ed.) *Annals of the New York Academy of Sciences Conference on Native and Foreign Language Acquisition*. New York: New York Academy of Sciences, pp. 190–214.

Boersma P (2001) Praat, a system for doing phonetics by computer. *Glot International* 5: 341–45.

Bordag D (2004) Interaction of L1 and L2 systems at the level of grammatical encoding: Evidence from picture naming. *EUROSLA Yearbook* 4: 203–30.

Bordag D and Pechmann T (2007) Factors influencing L2 gender processing. *Bilingualism: Language and Cognition* 10: 299–314.

Brouwer S, Sprenger S, and Unsworth S (2017) Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology* 159: 50–65.

Bruhn de Garavito J and White L (2003) The L2 acquisition of Spanish DPs: The status of grammatical features. In: Pérez-Leroux AT and Liceras J (eds) *The acquisition of Spanish morphosyntax: The L1/L2 connection*. Dordrecht: Kluwer, pp. 151–76.

Corbett G (1991) *Gender*. Cambridge: Cambridge University Press.

Dekydtspotter L, Schwartz BD, and Sprouse RA (2006) The comparative fallacy in L2 processing research. In: O'Brien MG, Shea C, and Archibald J (eds) *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006): The Banff Conference*. Somerville, MA: Cascadilla Press, pp. 33–40.

Dussias PE, Valdés Kroff JR, Guzzardo Tamargo RE, and Gerfen C (2013) When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition* 35: 353–87.

Faarlund JT, Lie S, and Vannebo KI (1997) *Norsk referansegrammatikk* [A reference grammar of Norwegian]. Oslo, Norway: Universitetsforlaget.

Franceschina F (2005) *Fossilized second language grammars: The acquisition of grammatical gender*. Amsterdam, Netherlands / Philadelphia, PA: John Benjamins.

Fretheim T (1985 [1976]) Er bokmålet tvekjønnet eller trekjønnet? [Does Bokmål have two or three genders?] In: Jahr EH and Lorentz O (eds) *Morfologi [Morphology]*. Oslo, Norway: Novus, pp. 99–101.

Göksel A and Kerslake C (2005) *Turkish: A comprehensive grammar*. London: Routledge.

Grüter T, Lew-Williams C, and Fernald A (2012) Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28: 191–215.

Guevara ER (2010) NoWaC: A large web-based corpus for Norwegian. In: Kilgarriff A, Lin D, and Sharoff S (eds) *Proceedings of the NAACL HLT 2010 6th web as corpus workshop*. Stroudsburg, PA: Association for Computational Linguistics, pp. 1–7.

Hawkins R and Chan CYH (1997) The partial availability of Universal Grammar in second language acquisition: The 'failed functional features hypothesis'. *Second Language Research* 13: 187–226.

Hopp H (2013) Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research* 29: 33–56.

Hopp H (2016) Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research* 32: 277–307.

Hopp H and Lemmerth N (2018) Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition* 40: 171–99.

Hovdenak M (1998) *Nynorskordboka: Definisjons- og rettskrivingsordbok [The new Norwegian dictionary: Dictionary of definitions and spellings]*. Oslo: Det Norske Samlaget.

Hulk A (2017) Note on cross-linguistic influence: Back to MULK. In: Blom E, Cornips L, and Schaeffer J (eds) *Cross-linguistic Influence in Bilingualism: In honor of Aafke Hulk*. Amsterdam, Netherlands: John Benjamins, pp. 15–24.

Kaan E and Grüter T (2021) Prediction in second language processing and learning: Advances and directions. In: Kaan E and Grüter T (eds) *Prediction in second language processing and learning*. Amsterdam, Netherlands: John Benjamins, pp. 2–24.

Klassen R (2016) The representation of asymmetric grammatical gender systems in the bilingual mental lexicon. *Probus* 28: 9–28.

Lardiere D (2009) Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research* 25(2): 173–227.

Lemhöfer K, Spalek K, and Schriefers H (2008) Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language* 59: 312–30.

Lenth RV (2021) *emmeans: Estimated marginal means, aka least-squares means: R Package Version 1.5.5-1*. Available at: https://CRAN.R-project.org/package=emmeans (accessed January 2024).

Lewis GL (2000 [1967]) *Turkish Grammar*. Oxford: Oxford University Press.

Lødrup H (2011) Hvor mange genus er det i Oslo-dialekten? [How many genders are there in the Oslo dialect?] *Maal og Minne* 2: 120–36.

Loerts H (2012) *Uncommon gender: Eyes and brains, native and second language learners, and grammatical gender*. Unpublished PhD dissertation, Rijksuniversiteit Groningen, Grodil, Netherlands.

Lundquist B and Vangsnes ØA (2018) Language separation in bidialectal speakers: Evidence from eye tracking. *Frontiers in Psychology* 9: 1394.

MacWhinney B (2008) A unified model. In: Ellis NC and Robinson P (eds) *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York: Erlbaum, pp. 341–72.

MacWhinney B and Bates E (1989) *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

Mirman D, Dixon JA, and Magnuson JS (2008) Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59: 475–94.

Montrul S, Foote R, and Perpinán S (2008) Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning* 58: 503–53.

Morales L, Paolieri D, and Bajo T (2011) Grammatical gender inhibition in bilinguals. *Frontiers in Psychology* 2: 284.

Morales L, Paolieri D, Dussias PE et al. (2016). The gender congruency effect during bilingual spoken-word recognition. *Bilingualism: Language and Cognition* 19: 294–310.

Paolieri D, Cubelli R, Macizo P et al. (2010). Grammatical gender processing in Italian and Spanish bilinguals. *Quarterly Journal of Experimental Psychology* 63: 1631–45.

Ragnhildstveit S (2017) *Genus og transfer når norsk er andrespråk: Tre korpusbaserte studier [Gender and transfer in Norwegian as a second language: Three* corpus-based studies.]. Unpublished PhD dissertation, University of Bergen, Bergen, Norway.

Rodina Y and Westergaard M (2015) Grammatical gender in Norwegian: Language acquisition and language change. *Journal of Germanic Linguistics* 27: 145–87.

Rodina Y and Westergaard M (2017) Grammatical gender in bilingual Norwegian–Russian acquisition: The role of input and transparency. *Bilingualism: Language and Cognition* 20: 197–214.

Rodina Y and Westergaard M (2021) Grammatical gender and declension class in language change: A study of the loss of feminine in Norwegian. *Journal of Germanic Linguistics* 33: 235–63.

Sabourin L and Stowe LA (2008) Second language processing: When are first and second languages processed similarly. *Second Language Research* 24: 397–430.
Sabourin L, Stowe LA, and de Haan GJ (2006) Transfer effects in learning a second language grammatical gender system. *Second Language Research* 22: 1–29.
Salamoura A and Williams JN (2007) The representation of grammatical gender in the bilingual lexicon: Evidence from Greek and German. *Bilingualism: Language and Cognition* 10: 257–75.
Singmann H, Bolker B, Westfall J, and Aust F (2016) *Afex: Analysis of factorial experiments: R Package Version 0.16–1*. Available at: https://CRAN.R-project.org/package=afex (accessed January 2024).
Stone K, Lago S, and Schad D (2021) Divergence point analyses of visual world data: Applications to bilingual research. *Bilingualism: Language and Cognition* 24: 833–41.
Trosterud T (2001) Genus i norsk er regelstyrt [Gender in Norwegian is rule-based]. *Norsk Lingvistisk Tidsskrift* 19: 29–58.
Tsimpli I, and Dimitrakopoulou, M (2007) The Interpretability Hypothesis: evidence from wh-interrogatives in second language acquisition. *Second Language Research* 23(2): 215–242.
Weber A and Paris G (2004) The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening. In: Forbus K, Gentner D, and Regier T (eds) *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, pp. 1446–51.

Appendix 1

Norwegian language proficiency test designed at UiT the Arctic University of Norway.

1. Hvor kommer dere fra?

☐ De kommer fra Spania
☐ Dere kommer fra Spania
☐ Vi er fra Spania
☐ De er fra Spania

2. Hvordan går det?
☐ Med buss
☐ Bare bra
☐ Jeg går på ski.
☐ Det går klokka 8.

3. Jeg elsker . . . mat.
☐ laget
☐ lage
☐ lager
☐ å lage

4. Unnskyld, kan du . . . meg litt?
☐ hjelper
☐ hjelpe
☐ å hjelpe
☐ hjelp

5. Vil du . . . litt kaffe?

☐ ha
☐ få
☐ lyst på
☐ med meg

6. Vi trenger åtte . . .
☐ stolen.
☐ stoler.
☐ en stol.
☐ stolene.

7. Når kommer han tilbake?
☐ I morgen.
☐ I morges.
☐ Sist fredag.
☐ I går.

8. Han er litt . . . gammel for meg.
☐ for
☐ til
☐ på
☐ to

9. . . . dag er det i dag?
☐ Hvilken
☐ Hva
☐ Hvordan
☐ Hvilke

10. Jeg vet ikke . . .
☐ om.
☐ når biblioteket åpner.
☐ dem.
☐ hvor bor han.

11. I dag . . .
☐ det er mandag.
☐ er mandag.
☐ må jeg være hjemme.
☐ jeg er syk.

12. Vi reiser til India . . .
☐ om to dager.
☐ i to dager.

☐ for to dager siden.
☐ før to dager.

13. Når . . . du å stå opp?
☐ skal
☐ pleier
☐ vekker
☐ vil

14. . . . du å hjelpe meg litt?
☐ Huske
☐ Har lyst på
☐ Kunne
☐ Orker

15. Hvor skal vi . . . kjøleskapet?
☐ sette
☐ ligge
☐ stå
☐ sitte

16. Hun vil sitte . . . siden av ham.
☐ i
☐ på
☐ ved
☐ til

17. Hvor er brillene?
☐ Denne ligge i bokhylla.
☐ Det ligger i bokhylla.
☐ Den ligger i bokhylla.
☐ De ligger i bokhylla.

18. Hva heter . . . Anna?
☐ mora av
☐ mor fra
☐ mora til
☐ mor av

19. Hun har en bror . . . bor i Oslo.
☐ som
☐ hvem
☐ at
☐ hvis

20. Det er fest . . . Ole i helga.

☐ hos
☐ til
☐ på
☐ i

21. Jeg vil ha en . . . ballong.
☐ gult
☐ gull
☐ gule
☐ gul

22. Dusjer du alltid . . .
☐ i morges?
☐ i morgen?
☐ om morgenen?
☐ i morgen tidlig?

23. Jeg gleder . . . til å se deg igjen.
☐ min
☐ seg
☐ deg
☐ meg

24. Jeg . . . sint hvis du gjør det igjen.
☐ er
☐ blir
☐ synes
☐ skal

25. Du spiser alltid så . . .
☐ rask.
☐ raskt.
☐ raske.
☐ raskere.

26. Hva . . . du i går?
☐ gjør
☐ gjøre
☐ gjorde
☐ har gjort

27. Jeg . . . så vondt i hodet.

☐ var
☐ gjorde
☐ følte
☐ hadde

28. Drømmer du . . . å bli advokat?

☐ av
☐ om
☐ at
☐ som

29. Du er mye flinkere . . . meg.

☐ enn
☐ som
☐ da
☐ den

30. Jeg skal ikke på jobb i dag.

☐ Jeg heller.
☐ Jeg også.
☐ Ikke jeg heller.
☐ Ikke jeg også.

31. Han sa at han . . . likte maten.

☐ har
☐ ikke
☐ fordi
☐ hvorfor

32. Da han kom, . . . for å spise.

☐ etterpå
☐ vi dro
☐ dro vi
☐ ville han

33. Hun . . . kastet ut.

☐ har
☐ ble
☐ snart
☐ ville

34. Jeg er . . . hjemme.

☐ mens
☐ midt i
☐ som regel
☐ de eneste

35. Hun prøvde . . . ham.

☐ å skylde
☐ å forsvinne
☐ å unngå
☐ å spandere

36. Hun likte den . . . bilen.

☐ hennes
☐ sin
☐ nye
☐ ny

Appendix 2

Experimental items.

		Gender		
Norwegian noun	English translation	Norwegian	Greek	Russian
bil	car	M	N	M
sko	shoe	M	N	M
sykkel	bicycle	M	N	M
knapp	button	M	N	F
ring	ring	M	N	M
blomst	flower	M	N	M
kjole	dress	M	N	N
nøkkel	key	M	N	M
hatt	hat	M	N	F
kniv	knife	M	N	M
sopp	mushroom	M	N	M
løk	onion	M	N	M
hanske	glove	M	N	F
hammer	hammer	M	N	M
vannmelon	watermelon	M	N	M
hals	neck	M	M	F
datamaskin	computer	M	M	M
vask	sink	M	M	F
foss	waterfall	M	M	M
drage	kite	M	M	M
passer	compass	M	M	M

(Continued)

Appendix 2. (Continued)

		Gender		
Norwegian noun	English translation	Norwegian	Greek	Russian
lighter	lighter	M	M	F
lysbryter	light switch	M	M	M
vannkoker	kettle	M	M	M
speil	mirror	N	M	N
gjerde	fence	N	M	M
belte	belt	N	F	M
bål	fire	N	F	M
slips	tie	N	F	M
anker	anchor	N	F	M
hus	house	N	N	M
bord	table	N	N	M
vindu	window	N	N	N
blad	leaf	N	N	M
tog	train	N	N	M
glass	glass	N	N	M
tre	tree	N	N	N
egg	egg	N	N	N
brød	bread	N	N	M
eple	apple	N	N	N
kjøleskap	refrigerator	N	N	M
sverd	sword	N	N	M
strykejern	iron	N	N	M
buss	bus	M	N	n/a
sofa	couch	M	M	n/a
ovn	oven	M	M	n/a
rakett	rocket	M	M	n/a
satellitt	sattelite	M	M	n/a
albue	elbow	M	M	n/a
linjal	ruler	M	M	n/a
sirkel	circle	M	M	n/a
bein	bone	N	N	n/a
kamera	camera	N	F	n/a
tak	roof	N	F	n/a
fengsel	prison	N	F	n/a
kart	map	N	M	n/a
kjøkken	kitchen	N	F	n/a
kors	cross	N	M	n/a
basseng	swimming pool	N	F	n/a
stempel	rubber stamp	N	F	n/a
jordbær	strawberry	N	F	n/a
badekar	bathtub	N	F	n/a

(Continued)

Appendix 2. (Continued)

Norwegian noun	English translation	Gender		
		Norwegian	Greek	Russian
fjell	mountain	N	N	n/a
skjerf	scarf	N	N	n/a
vei	road	M	n/a	M
katt	cat	M	n/a	F
penn	pen	M	n/a	F
kurv	basket	M	n/a	F
gaffel	fork	M	n/a	F
ballong	balloon	M	n/a	M
brev	letter	N	n/a	N
hjerte	heart	N	n/a	N
(tre-)hull	tree hollow	N	n/a	N
øre	ear	N	n/a	N
egg	egg	N	n/a	N
hjul	wheel	N	n/a	N
korn	grain	N	n/a	N
spøkelse	ghost	N	n/a	N
håndkle	towel	N	n/a	N
nordlys	aurora borealis	N	n/a	N
gevær	rifle	N	n/a	N
spyd	spear	N	n/a	N
eple	apple	N	n/a	N
brev	letter	N	n/a	N

Notes. F = feminine. M = masculine. N = neuter. n/a = not applicable.

Appendix 3

Results of the generalized linear mixed-effects model for gender assignment accuracy fitted to the data set of the L1 Greek and L1 Turkish participants. There were in total 2,381 datapoints, 43 participants and 64 items. The model contains random intercepts for Participant (variance: 0.32, *SD*: 0.566) and Item (variance: 0.53, *SD*: 0.728). The following model was fitted:

glmer(GenderAccuracy ~ (L2 Proficiency × L1 × Norwegian–Greek Lexical Gender Congruency) + (Norwegian Target Gender × L1 × L2 Proficiency) + (Norwegian Target Gender × Norwegian–Greek Lexical Gender Congruency) + (1|Part) + (1|Item), family = binomial).

Proficiency was centered at mean, and all binary variables were sum coded, where the estimates give the difference between the overall mean and the first level of each predictor, here Greek, Congruent and Masculine. The chi-square statistics reported in the text are given in the final two columns. These values were obtained from model comparisons using the afex package (Singmann et al., 2016).

	Estimate	SE	z-value	(df) Chisq	*p*
(Intercept)	1.42954	0.14230	10.046		
Prof	3.68732	0.94869	3.887	(1) 13.11	< .001***
L1	0.07390	0.10731	0.689	(1) 0.47	.493
Cong	0.36600	0.11077	3.304	(1) 10.21	.001**
NorGen	0.63390	0.11089	5.717	(1) 27.10	< .001***
Prof:L1	−1.24987	0.94394	−1.324	(1) 1.69	.194
Prof:cong	0.07500	0.54225	0.138	(1) 0.02	.890
L1:cong	0.07960	0.05794	1.374	(1) 1.84	.174
Prof:NorGen	−2.68615	0.55398	−4.849	(1) 24.07	< .001***
L1:NorGen	0.10721	0.05899	1.817	(1) 3.24	.072
cong:NorGen	−0.21287	0.11016	−1.932	(1) 3.58	.058
Prof:L1:cong	−0.10666	0.53440	−0.200	(1) 0.04	.843
Prof:L1:NorGen	−0.03941	0.54915	−0.072	(1) 0.01	.943

Notes. **$p < .01$. ***$p < .001$.

Appendix 4

Results of the generalized linear mixed-effects model for gender assignment accuracy fitted to the data set of the L1 Russian participants. There were in total 1,220 datapoints, 23 participants and 64 items. The model contained random intercepts for Participant (variance: 0.437, *SD*: 0.661) and Item (variance: 0.913, *SD*: 0.955). The following model was fitted:

glmer(GenderAccuracy ~ (L2 Proficiency × Norwegian–Russian Lexical Gender Congruency × Norwegian Target Gender) + (1|Part) + (1|Item), family = binomial).

L2 Proficiency was centered at mean, and all binary variables were sum coded, where the estimates give the difference between the overall mean and the first level of each predictor, here Congruent and Masculine. The chi-square statistics reported in the text are given in the final two columns. These values are obtained from model comparisons using the afex package (Singmann et al., 2016).

	Estimate	SE	z-value	(df) Chisq	*p*
(Intercept)	1.266	0.201	6.31		
Prof	3.505	1.076	3.26	(1) 9.04	.003**
Cong	−0.117	0.140	−0.84	(1) 0.69	.406
NorGen	0.592	0.143	4.13	(1) 16.07	< .001***
Prof:Cong	−0.187	0.517	−0.36	(1) 0.13	.722
Prof:NorGen	0.944	0.519	1.82	(1) 3.23	.072
cong1:NorGen	0.038	0.138	0.27	(1) 0.07	.785
Prof:Cong:NorGen	−0.150	0.517	−0.29	(1) 0.08	.775

Notes. **$p < .01$. ***$p < .001$.

Appendix 5

Coefficients from the generalized linear mixed-effects model on gender processing for the early RoI, with the predictors Cond(ition), (L2) Prof(iciency), Lang(uage) and (Norwegian Target) Gender, and all interactions involving these variables (up to 4 levels). L2 Proficiency was centered at mean, and the categorical variables were sum coded (Cond1 = Different, Lang1 = Turkish, Lang2 = Greek, Gender1 = Neuter). The model contained random intercepts for Item (Variance: 0.0277, *SD*: 0.166) and Participant (Variance: 0.3754, *SD*: 0.613) and a by-participant slope for Condition (Variance: 0.0104, *SD*: 0.102). There were 7,610 observations, 85 items and 66 participants.

	Estimate	SE	z-value	Pr (> | t |)
(Intercept)	−0.415898	0.083517	−4.98	6.4e–07***
cond1	0.050642	0.028445	1.78	0.07502
Prof	−0.009827	0.417939	−0.02	0.98124
Lang1	0.000684	0.117848	0.01	0.99537
Lang2	−0.261850	0.114923	−2.28	0.02270*
gender1	−0.001147	0.031359	−0.04	0.97083
cond1:Prof	0.526342	0.147932	3.56	0.00037***
cond1:Lang1	−0.027067	0.040305	−0.67	0.50186
cond1:Lang2	0.019094	0.040634	0.47	0.63842
Prof:Lang1	−1.049435	0.549656	−1.91	0.05623
Prof:Lang2	0.181435	0.617273	0.29	0.76881
cond1:gender1	−0.015258	0.024993	−0.61	0.54152
Prof:gender1	−0.007295	0.130512	−0.06	0.95542
Lang1:gender1	0.003622	0.036000	0.10	0.91986
Lang2:gender1	−0.038745	0.036500	−1.06	0.28845
cond1:Prof:Lang1	−0.493814	0.191065	−2.58	0.00975**
cond1:Prof:Lang2	0.232037	0.225049	1.03	0.30252
cond1:Prof:gender1	−0.103601	0.130483	−0.79	0.42721
cond1:Lang1:gender1	−0.016424	0.035618	−0.46	0.64472
cond1:Lang2:gender1	−0.024300	0.036089	−0.67	0.50073
Prof:Lang1:gender1	0.146081	0.167672	0.87	0.38363
Prof:Lang2:gender1	−0.284059	0.197773	−1.44	0.15092
cond1:Prof:Lang1:gender1	0.027259	0.167656	0.16	0.87084
cond1:Prof:Lang2:gender1	0.010969	0.197702	0.06	0.95575

Notes. *$p < .05$. **$p < .01$. ***$p < .001$.

Appendix 6

Coefficients from the generalized linear mixed-effects model on gender processing for the late RoI, with the predictors Cond(ition), (L2) Prof(iciency), Lang(uage) and (Norwegian Target) Gender, and all interactions involving these variables (up to 4 levels). Proficiency was centered at mean, and the categorical variables were sum coded (Cond1 = Different, Lang1 = Turkish, Lang2 = Greek, Gender1 = Neuter). The model contained random intercepts for Item (Variance: 0.0193, *SD*: 0.139) and Participant (Variance: 0.3736, *SD*: 0.611) and a by-participant slope for Condition (Variance: 0.0177, *SD*: 0.133). There were 7,902 observations, 85 items and 66 participants.

	Estimate	SE	z-value	Pr (> | t |)
(Intercept)	−0.22602	0.08240	−2.74	0.0061**
cond1	0.12268	0.02968	4.13	3.6e–05***
Prof	0.65639	0.41779	1.57	0.1162
Lang1	0.02650	0.11716	0.23	0.8210
Lang2	−0.22916	0.11411	−2.01	0.0446*
gender1	0.04776	0.02896	1.65	0.0991
cond1:Prof	0.91575	0.15762	5.81	6.3e–09***
cond1:Lang1	−0.03102	0.04239	−0.73	0.4643
cond1:Lang2	0.05814	0.04225	1.38	0.1688
Prof:Lang1	−1.51258	0.54795	−2.76	0.0058**
Prof:Lang2	0.90717	0.61733	1.47	0.1417
cond1:gender1	−0.00634	0.02421	−0.26	0.7936
Prof:gender1	0.02299	0.12833	0.18	0.8578
Lang1:gender1	0.03903	0.03486	1.12	0.2630
Lang2:gender1	−0.00596	0.03494	−0.17	0.8646
cond1:Prof:Lang1	−0.52247	0.20301	−2.57	0.0101*
cond1:Prof:Lang2	0.29548	0.23744	1.24	0.2133
cond1:Prof:gender1	−0.21065	0.12834	−1.64	0.1007
cond1:Lang1:gender1	0.04714	0.03455	1.36	0.1725
cond1:Lang2:gender1	0.01334	0.03461	0.39	0.7000
Prof:Lang1:gender1	−0.11569	0.16463	−0.70	0.4822
Prof:Lang2:gender1	0.13481	0.19248	0.70	0.4837
cond1:Prof:Lang1:gender1	0.11942	0.16464	0.73	0.4682
cond1:Prof:Lang2:gender1	−0.08337	0.19250	−0.43	0.6649

Notes. *$p < .05$. **$p < .01$. ***$p < .001$.

Appendix 7

Lexical gender congruency effects. The table summarizes the chi-square statistics for the six sets of models, one for each language in each regions of interest (RoI), as obtained from the afex package (Singmann et al., 2016).

	df	Greek χ^2	Greek *p*	Turkish χ^2	Turkish *p*	Russian χ^2	Russian *p*
Early RoI:							
Condition	1	2.01	.156	0.48	.490	1.36	.244
Prof2_c	1	0.09	.767	3.99	.046*	1.69	.193
Congruency	1	0.08	.777	0.27	.601	1.26	.261
Condition:Prof2_c	1	3.52	.061	0.09	.761	7.74	.005**
Condition:Congruency	1	1.84	.175	0.49	.483	0.54	.464
Prof2_c:cong	1	5.27	.022*	0.16	.690	1.08	.299
Condition:Prof2_c:cong	1	0.97	.325	0.45	.504	2.90	.089
Late RoI:							
Condition	1	10.16	.001**	2.21	.137	3.94	.047*
Prof2_c	1	2.74	.098	2.69	.101	3.50	.061
Congruency	1	0.03	.858	0.32	.570	0.23	.630
Condition:Prof2_c	1	10.93	< .001***	2.33	.127	13.04	< .001***
Condition:Congruency	1	0.15	.697	1.42	.233	0.99	.320
Prof2_c:cong	1	0.57	.449	0.08	.774	1.22	.270
Condition:Prof2_c:cong	1	2.89	.089	2.25	.133	0.01	.938

Notes. *$p < .05$. **$p < .01$. ***$p < .001$.

Appendix 8

Grouping of data. In the model presented in Table 5, we include all the three language groups, and in models 6–8 we, furthermore, present the coefficients from models for the three individual groups. As pointed out by reviewers, it also makes sense to look at the Greek and Turkish group together, excluding the Russian group, in line with the analysis of the production data. Below we present the coefficients from models including only Greek and Turkish, from both early and late RoI. Gender (Masculine/Neuter) is not included as a fixed predictor in the models to make the results more easily interpretable. Gender is included as a random intercept in the model for the late RoI, but had to be excluded from the early model due to problems with singularity of the model. The models have the following structure:

$$\text{TargetLook} \sim \text{cond} \times \text{Lang} \times \text{Prof2\_c} + (1 + \text{cond}|\text{Participant}) + (1|\text{Item}) + (1|\text{gender}).$$

There were 4,969 observations in the Early model, and 5,177 observations in the late model. In both the early and the late model, we found a significant three-way interaction between condition, proficiency and language, similarly to the model fitted for the three languages. For a more detailed

comparison between the Greek and Turkish groups throughout the whole trials, see Appendix 9, Appendix Figure 3.

	Estimate	SE	*z*-value	Pr (> | t |)
Early:				
(Intercept)	−0.75197	0.14947	−5.031	4.89e–07***
condDiff	0.15344	0.09994	1.535	0.12469
LangTurk	0.31408	0.21484	1.462	0.14376
Prof2_c	−0.54285	0.83059	−0.654	0.51339
condDiff:LangTurk	−0.10179	0.13849	−0.735	0.46235
condDiff:Prof2_c	1.48719	0.57665	2.579	0.00991**
LangTurk:Prof2_c	−0.55211	1.05396	−0.524	0.60038
condDiffLangTurk:Prof2_c	−1.40837	0.70674	−1.993	0.04629*
Late:				
(Intercept)	−0.6337	0.1501	−4.223	2.41e–05***
condDiff	0.3597	0.1056	3.406	0.000659***
LangTurk	0.3473	0.2093	1.659	0.097055
Prof2_c	0.3848	0.8139	0.473	0.636383
condDiff:LangTurk	−0.1765	0.1490	−1.185	0.236190
condDiff:Prof2_c	2.4702	0.6287	3.929	8.53e–05***
LangTurk:Prof2_c	−1.6391	1.0320	−1.588	0.112207
condDiff:LangTurk:Prof2_c	−1.6823	0.7714	−2.181	0.029183*

Notes. *$p < .05$. **$p < .01$. ***$p < .001$.

Appendix 9

The reviewers of the article provided several suggestions for alternative analyses of the eye-tracking data, of which some were included in the final analysis. In this appendix, we discuss in more detail why we made certain choices and also present alternative analyses. There are mainly three points where our analysis may appear non-standard, which we list below:

1. Including all trials in the analysis, independent of the participants' performance for the target and distractor items in the preceding production task.
2. Including looks to white space in the analysis, i.e. fixations that were neither on the target nor on the distractor.
3. Defining the regions of interest based on the control groups' gaze patterns.

We discuss these three points in detail below. Regarding the first point, it might appear obvious to exclude items that participants could not name or to which they could not assign the correct gender: How could participants make predictions based on grammatical gender if they do not know the gender in the first place? However, there are several reasons to include these items. First, it is a simplification to equate the individual results from the production test with 'knowledge' of gender.

It presupposes a simplistic binary understanding of grammatical knowledge; either you know it or you do not. In many cases, this is probably not the case. In the case of Norwegian, L2 speakers are most likely uncertain about the gender of many nouns. In the production task, L2 speakers have to allocate most effort on retrieving the correct noun and spend less effort getting the gender right, which may result in participants defaulting to the masculine gender. The performance on the neuter nouns was low, but it is likely that accuracy would have been higher in a perception task compared to a production task, e.g. if the participants had been presented with each noun preceded by both the neuter and masculine articles, and had been required to choose which one sounded correct, i.e. in a context where the participant did not have to retrieve the correct noun and, also, in a context where they heard the word produced by a native speaker. A related problem with excluding items based on gender 'knowledge' is the default status of the masculine. It is impossible to say if a speaker 'knows' that a noun is masculine, or if they just use the default article *en*. Still, we believe that the overall scores from the production experiment tell us (a) which nouns are difficult to assign gender to, and (b) how proficient the individual participants are with respect to grammatical gender. However, it is not possible to tell whether an individual participant knows the gender of a particular noun or not.

Another obvious problem of removing items based on the production results is the loss of data. Every noun/image is used four times in the experiment: as a target and competitor in both the same-gender and different-gender conditions. This means that we would have to remove four eye-tracking items for every production error. For a participant who scores 75% correct on the production test, we would have to remove as much as 50% of the eye-tracking data (possibly less, depending on the distribution of the errors). Even more problematic, the eye-tracking data would be highly unbalanced after removal due to the higher proportion of errors for neuter (and incongruent) nouns. We would be able to keep most of the same-gender masculine trials, a decent amount of different-gender masculine and neuter trials, but very few same-gender neuter trials (the same problem would arise for congruency). For some neuter nouns, we would have no trials for the lower-proficiency speakers. This would make comparisons across proficiency groups unreliable, as the groups are tested on partially different sets of nouns; more specifically, the high-proficient speakers are tested on a larger and presumably more difficult set of nouns. Furthermore, since every noun appears four times in the experiment, it is not implausible that the speakers 'learn' the gender of the noun after its first appearance. Would it then make sense to remove only the first occurrence of the word? Probably not, but this would be another point one would have to decide on. Finally, it should be pointed out that the 'same/different competitor gender' paradigm used in this article is relatively resistant to the problem of incorrect gender knowledge. Associating a noun with the incorrect gender will not lead to a reversal of the same/different effect: it should only neutralize the same/different manipulation. Thus, even for participants with a fairly high amount of gender errors, it should still be possible to see an effect of the manipulation, driven solely by the cases of correct gender assignment (if these participants indeed engage in predictive gender processing).

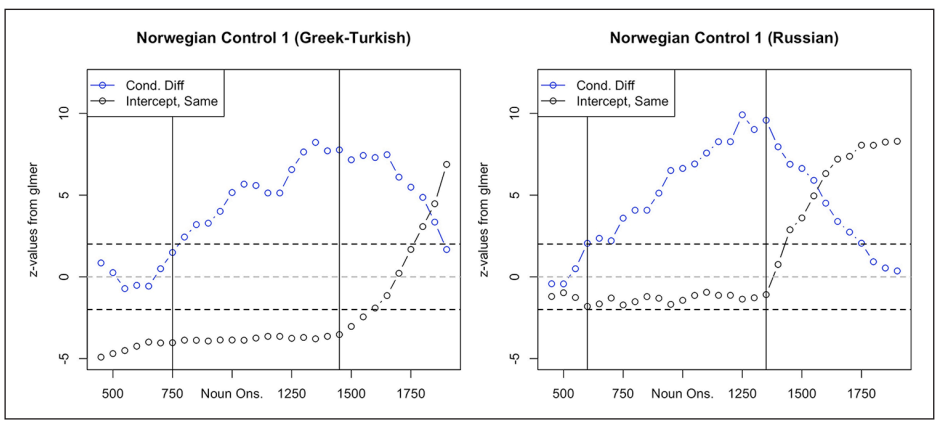
As for the second point, there is some variation in the eye-tracking literature with respect to the inclusion or exclusion of fixations outside the target and distractor images in the analysis. Excluding looks to white space may facilitate analysis, especially in a two-picture eye-tracking paradigm where only two regions of interest remain after white-space removal, which helps make the intercept and coefficients from logistic regression models more interpretable. Yet, at the onset

of each trial, there are three, not two, types of behaviors that can be observed: the participant may saccade to the left, to the right, or keep the fixation at the middle of the screen. Participants vary to a high degree in their behavior. Many participants choose to stay in the middle of the screen until some linguistic disambiguating information has been presented. By excluding the data related to the third option, i.e. fixating white space, the researcher basically rejects looks to white space as a behavioral response. This is potentially problematic, as a lot of highly meaningful data is thrown away. Looks to whitespace after article onset in a same-gender trial means something different than looks to whitespace in a different-gender trial, and the difference is equally meaningful as fixations to the distractor in same- and different-gender trials. Removing looks to white space is thus both a waste of data and potentially a source of misinterpretation of the data.

The final point to justify is the choice of the regions of interest, i.e. the early and the late regions of interest, as used in the analysis. Our choice is perhaps not the ideal one, and an analysis that could reflect the participants' behavior across the whole trial rather than just at two 50 ms-regions would be preferable. Yet, given the complexity of the data sets and the research questions, any analysis that takes into account the development of fixations over time – like a time series analysis – would have to introduce one more level of interactions, ending up with five or six levels of interactions. Our data is not large enough for this and, if it were, the model outputs would be very hard to interpret. Another option would be to take time of first fixation of the target image as the dependent variable. This is however problematic for a two-picture paradigm with a relatively long time between presentation of image stimuli and onset of the gender-carrying article: participants have often fixated the target image prior to the onset of the article. These trials with early target fixation would simply have to be excluded from the analysis. Divergence point analyses are also problematic here as proficiency is coded as a continuous value. Yet, the seemingly arbitrary choices of regions of interest in this article may lead the reader to suspect that we have cherry-picked the regions where we find differences between groups. This is not the case: the regions of interests are based firmly on the behavior patterns found in the control group data. There are many factors influencing the temporal development of fixations in a visual world paradigm gender study, e.g. number of pictures on the screen, participants' familiarity with the material, complexity of the gender system, and the particular phonetic properties of the articles. In Norwegian, both the masculine and neuter articles start with a mid-open, central vowel. It is not clear if native speakers can pick out the differences between the two articles based on the vowel itself, or if the point of disambiguation takes place at the coda. The same holds for the onset of the noun. In our study, we found that the fixation patterns differed between the two experiments: looks to target (both in the same- and different-gender conditions) increased earlier in the Russian data set compared to the Greek/Turkish data set, as was seen both in the control groups and the test groups. It is unclear whether this may be due to differences in the material, or due to the hardware (small screen laptop vs. larger monitor and different eye-trackers). The only sensible approach here is to define regions of interest based on the control groups' behavior on the same stimuli, and not on saccade and fixation patterns from studies carried out on other participant groups with different stimuli and hardware.

Figures 5 to 7 indicate that the differences between the groups are visible for a larger stretch of time than just the two regions of interests. Yet, those figures do not by themselves show that there are significant differences between the three L1 groups; only a model that includes all three groups can do that. Figures 5 to 7 do not illustrate the continuous proficiency measure used in the models

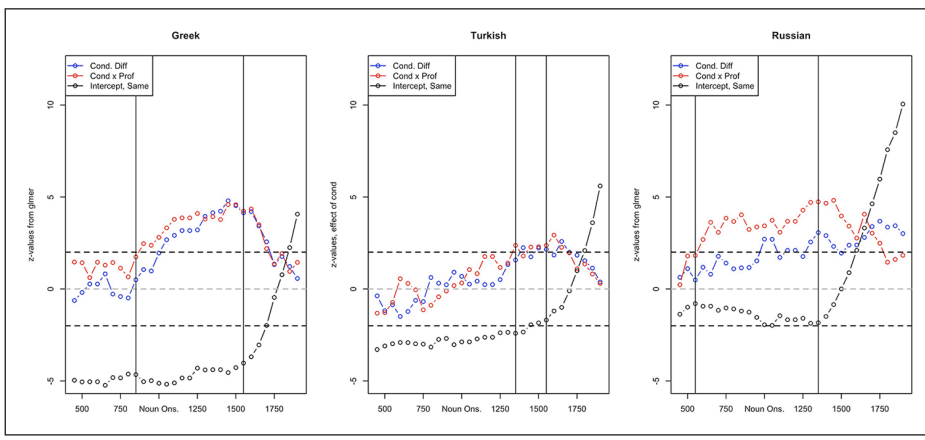
either. To address these issues, we provide graphs of the time course which include the continuous proficiency measures; see Appendix Figures 1–4. In these figures, we plot not the proportions of looks, but the *z*-values related to coefficients for condition and the interaction between proficiency and condition for the different groups, and pairwise compare the three groups. The *z*-values come from the mixed-effects logistic regression models with L1, proficiency, and condition as predictors, as well as the interactions between these predictors, and with participant and item as random intercepts. We fitted one model per time slot, and changed the reference level of the intercept to each L1 group. First, we look at the control groups and L1 groups separately. In the graphs in Appendix Figure 1, we show the *z*-values related to the intercept, here, looks to the target in the same-gender condition, and the coefficient for condition (different-gender) for the two control groups. Note that the graphs from the control group present the same data as in Figure 4. The dark line shows the intercept, here, looks to the target in the same-gender condition, which is below zero, i.e. below chance, as fixations are distributed across the target, distractor, and white space. A couple of hundred milliseconds after the noun onset, the intercept rapidly increases, which indicates that the fixations are guided by the noun itself, and not the gender value. As was discussed in connection with Figure 4, this effect is earlier in the Russian control group by around 100 ms. The blue line shows the effect of condition, i.e. the increase in looks to the target when there is only one referent on the screen that matches the gender expressed by the article. Also here, the effect of the gender value is shown earlier in the Russian data set, already at around 500–600 ms, compared to the Greek/Turkish data set. The horizontal lines at −2 and 2 represent an approximate 'significant' result at an alpha level of 0.05 (the value of the coefficient is twice as big as the standard error). The two vertical lines indicate the limits of the time region where the participants' fixations are guided by the gender marked article (i.e. between the point where the red line approaches a *z*-value of 2, and the inflection point of the black line). For both control groups, we see the effect for about 700 ms. Note that there is about 1,000 ms from the onset of the article to the onset of the noun, which suggests that it takes about 300 ms extra for participants to respond to the information coded in the article compared to the noun. If this difference is due to shared onset of the two articles or higher processing load associated with functional compared to lexical cues, is still unknown.



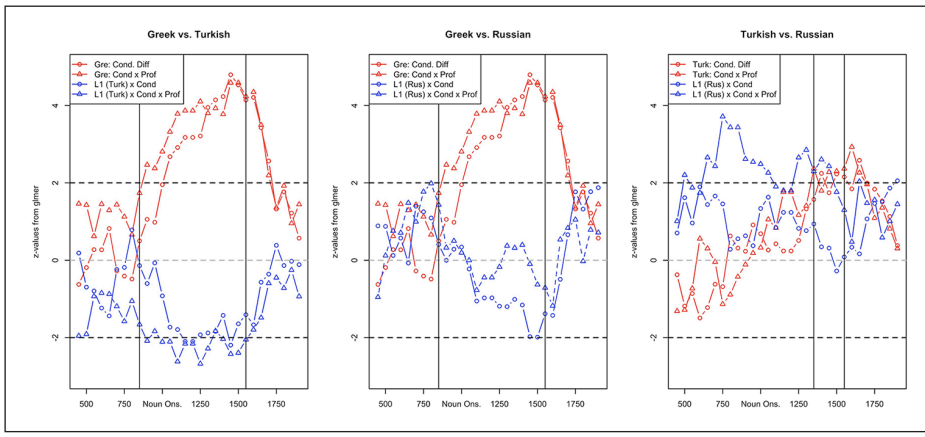
Appendix Figure 1. *Z*-values from glmer for the two control groups.
Note. The black line is the intercept (Same gender) and the blue line is the effect of condition (same/diff).

To illustrate the effect of condition in the L2 groups, we include the interaction effect of proficiency and condition in addition to the main effect of condition. As proficiency is centered at mean, the *z*-value of condition (same/different) gives the estimated effect of condition (blue line) at mean proficiency, while the interaction effect (red line) gives the increase in effect as proficiency increases. As can be seen, there is a salient effect of condition in the Greek group from about noun onset. The interaction between condition and proficiency is salient more or less in the same temporal region as the effect of condition is in Norwegian control group 1. The effect of the interaction between condition and proficiency in the Russian group also shows more or less the same temporal profile as the effect of condition in control group 2. The main effect of condition is less salient though. In the Turkish group, it is hard to detect a salient effect of condition or an interaction between condition and proficiency until about 400 ms after noun onset. In short, these graphs show a pattern similar to the ones in Figures 5 to 7.

Still, Appendix Figure 2 cannot show if there are significant group differences with respect to the same/different gender manipulation. In the three graphs below (Appendix Figure 3) we plot pairwise comparisons of the three language groups. In the left-hand figure, we compare Turkish to Greek. We see a significant two-way interaction between condition and L1 (blue line with dots) at about 1,100–1,500 ms after article onset. The three-way interaction (L1 × condition × proficiency) lasts for the whole period where the 'condition × proficiency' interaction is seen for Greek. When comparing Greek and Russian, we find no clear significant difference between the groups for a longer stretch of time, i.e. both blue lines mostly stay within ±2. The L1 Russian group shows an interaction effect of proficiency and condition earlier than the L1 Greek group, and the effect of condition is smaller at around 1,500 ms after article onset, which is expected given that the effect of condition both starts and ends earlier in the Russian group compared to the other two groups. The difference between the Turkish and Russian groups mainly comes out as a three-way interaction (L1 × condition × proficiency).

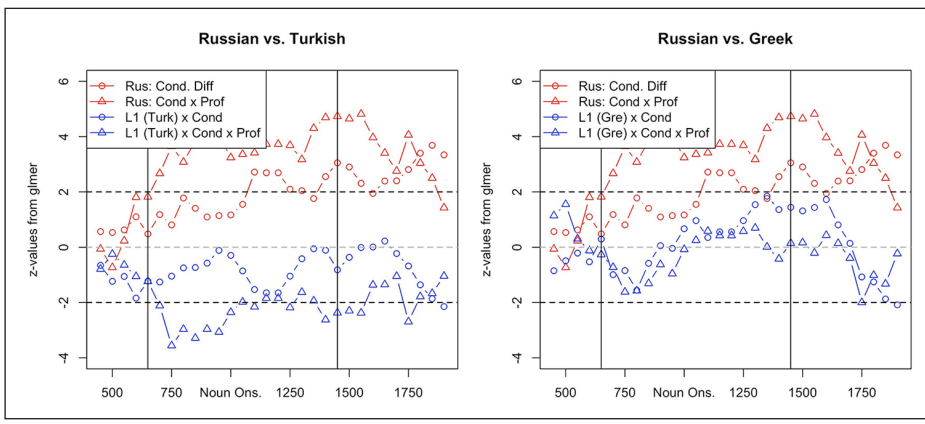


Appendix Figure 2. *Z*-values from glmer from the three first language (L1) groups.
Note. The black line is the intercept (Same gender), the blue line is the effect of condition and the red line is the interaction between condition and proficiency.



Appendix Figure 3. Pairwise comparisons between the languages. Greek is the intercept language in the left-hand and middle figures, and Turkish is the intercept in the right-hand figure.

A final concern about the comparisons between the L2 groups is the differences in the temporal development between the two control groups. The differences between the Russian and Turkish groups may be a result of everything taking place earlier in the Russian group data set. The difference between the Greek and Russian group may also be affected by this. As a final measure of precaution, we fit models for each time slot, with all the Russian measures shifted 100 ms back in time. We plot the pairwise comparisons in Appendix Figure 4, with Russian as the intercept group. As can be seen, the difference between the Russian and Turkish groups is still salient (mainly as a three-way interaction), while the Russian and Greek groups look even more similar with the shifted Russian measures.



Appendix Figure 4. Pairwise comparisons with Russian as the intercept, where the time stamps in the Russian data set has been shifted 100 ms, in order to account for the difference in temporal profiles between the two data sets.