Cognition 224 (2022) 105060

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

The mechanisms underlying grammatical gender selection in language production: A meta-analysis of the gender congruency effect

Ana Rita Sá-Leite [a,*], Karlos Luna [b], Ângela Tomaz [c], Isabel Fraga [a], Montserrat Comesaña [d,e]

[a] Cognitive Processes & Behaviour Research Group, Department of Social Psychology, Basic Psychology & Methodology, University of Santiago de Compostela, Spain
[b] Departamento de Psicología, Universidad Nacional de Colombia, Colombia
[c] Sciences Cognitives et Sciences Affectives (SCALab), University of Lille, France
[d] CIPsi, School of Psychology, University of Minho, Portugal
[e] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Spain

ARTICLE INFO

Keywords:
Gender agreement
Gender congruency effect
Grammatical gender
Picture-word interference paradigm

ABSTRACT

Grammatical gender retrieval during language production has been largely addressed through the picture-word interference (PWI) paradigm, with the aim of capturing the so-called gender congruency effect (GCE). In the PWI paradigm, participants name target pictures while ignoring superimposed written distractor nouns. The GCE shows faster responses when target and distractor nouns share the same gender than when gender differs. Yet, the locus of this effect is not clear: it might be either due to the selection of a determiner or due to the selection of a gender node at the lemma level, which may be primed or delayed by competition. Importantly, many of those who argue that the GCE is not a genuine effect of gender conclude that gender is a feature that is retrieved automatically. Such a claim is controversial since the PWI paradigm has been seen as too complex and perhaps not sensitive enough to capture small effects. Besides, for Romance languages, mixed results draw a complex picture with effects occurring mainly in the opposite direction, i.e., a gender incongruency effect (GIE). In the present study, we conducted a meta-analysis of the 18 studies that have addressed this issue. The results confirm the existence of the GCE as a determiner effect in Germanic/Slavic languages, while little support is found for the GIE in Romance languages. Nevertheless, we argue that the absence of gender effects in Germanic and Slavic languages within the PWI paradigm cannot be taken as evidence of an absence of priming/competition during gender selection and thus as evidence of an automatic selection of gender. Parametric replication of previous studies, especially those featuring bound morphemes, together with the use of other measuring techniques such as event related potentials are suggested as a way forward.

1. Introduction

The study of grammatical gender retrieval during language production has prompted heated debates regarding the mechanisms underlying gender selection (for an overview, see Wang & Schiller, 2019). Special attention has been given to the nature of these mechanisms, with some authors defending that they have a facilitative priming basis, others a competitive one, and others arguing that gender is selected automatically without facilitative priming or competition involved. A number of methodological aspects has also been a target of criticism and controversy, as results have been quite mixed and slippery, especially when considering different elements of agreement within the produced utterance across language families. Certain effects are said to speak against the existence of priming or competition during gender selection within specific paradigms, but as we will see, they are based on the absence of evidence rather than on measurable supporting evidence. Others simply do not fit the tenets of any of the classical models of language production. Importantly, interpretations based on these effects can have great repercussions in the way we represent grammatical gender and conceive the architecture of lexical access. In this study, we first define the main problematic aspects of the literature on this subject in both theoretical and methodological terms, and then present a meta-analysis that assesses the size of the main effects found in this area through the calculation of Hedges *g*. We aim to determine the most robust effects, identify those urgently calling for replication, and analyse the existence of potential publication bias. Finally, we propose different ways of

* Corresponding author at: Cognitive Processes & Behaviour Research Group, University of Santiago de Compostela, 15782, Spain.
E-mail address: anarita.saleite.dias@usc.es (A.R. Sá-Leite).

https://doi.org/10.1016/j.cognition.2022.105060
Received 2 October 2020; Received in revised form 4 February 2022; Accepted 6 February 2022
Available online 18 February 2022

overcoming the disputes previously identified, and stress the limitations of the task typically used within this subject: the picture-word interference (PWI) paradigm.

The PWI task was first used to explore the underlying mechanisms of gender selection in Dutch (Schriefers, 1993). More specifically, Schriefers asked participants to name a series of pictures that were superimposed with written distractor nouns (e.g., naming the picture of a chair, "*stoel*", while ignoring the word "castle", "*kasteel*"). To do so, participants had to use noun phrases formed by either a definite article, an adjective, and a noun (e.g., "*de groene stoel*" [the green chair]) or an adjective plus a noun (e.g., "*groene stoel*" [green chair]). Both the definite article and the adjective varied according to gender (e.g., "*groene stoel*" [common gender] vs. "*groen huis*" [neuter gender; "green house"]). The picture noun (i.e., the target) and the distractor noun could have either the same or opposite gender values. Results showed what was called the gender congruency effect (GCE): participants were faster when target and distractor nouns had the same gender (e.g., the common target "*stoel*" benefited more from being paired with a common distractor such as "*tafel*" ["table"] than with a neuter distractor such as "*kasteel*"; the same happened for neuter targets).

Schriefers (1993) interpreted the GCE according to the most widely accepted structure of lexical access (see Roelofs, 1992). It was mainly conceived as a process of spreading of activation occurring through different representational levels of conceptual, syntactic, and morpho-phonological information. On the syntactic stratum, grammatical features were thought to be represented as nodes that accumulated activation. In the case of gender, each gender node represented a specific value (e.g., neuter and common nodes in Dutch) and nouns of a certain gender value were linked to their respective gender node (e.g., "*stoel*" would be connected to the common node). The node reaching an absolute threshold of activation would be selected. Yet, according to Schriefers (1993), each gender value would compete for selection, thereby leading to the GCE. The most influential models of language production that have explicitly addressed gender in their tenets partially deviate from this interpretation. More specifically, the Word Encoding by Activation and VERification model, or WEAVER++ (Levelt, Roelofs, & Meyer, 1999), and the Independent Network model (IN, Caramazza, 1997), also view gender as nodes located on a syntactic stratum, but this stratum is, for the former model, a lemma mediating between semantics and morphophonology, whereas for the latter it is an independent syntactic-grammatical network (see Fig. 1). Importantly, both models, especially WEAVER++, included certain nuances that allowed for greater flexibility in terms of explaining experimental results. WEAVER++ stated that gender was always activated, but only selected when necessary to determine the form of other words. Thus, selection of a gender node only occurred in the presence of agreement elements (e.g., definite articles, in Dutch, "*de*" [common] vs. "*het*" [neuter]). Moreover, unlike what Schriefers (1993) proposed, WEAVER++ assumes that a gender node is selected when its activation exceeds an absolute activation threshold, without competition (see Roelofs, 2018).[1] Hence, the GCE is seen as an effect of facilitative priming rather than of competition. That is, in the congruent condition, the gender node of the target will be primed by the gender value of the distractor, and therefore will exceed the threshold quicker than in the incongruent condition. The IN model did not explicitly address the role of agreement, and claimed that, because of lexical selection, gender is always selected in a competitive process that ends with the inhibition of the non-target gender node. It also proposed an architecture of lexical access that entailed direct connections between the conceptual and morpho-phonological networks. Ultimately, this allows the production of a noun without the intervention of the grammatical syntactic network, thus making it conceivable that gender encoding may not always occur.

[1] We would like to thank an anonymous reviewer for raising this point.

After the observation of this initial clear-cut GCE, subsequent studies introduced heated debate to the field of gender processing and to lexical access in general, challenging the views of the abovementioned models. In short, they test two interpretations of the locus of the gender congruency effect: the effect is due to selection of a gender node at the lemma level (which may be primed [WEAVER++, Levelt et al., 1999] or delayed by competition [IN model, Caramazza, 1997]), or due to the selection of an element of agreement. Studies with Germanic and Slavic languages replicated Schriefers' results when considering noun phrases formed by either definite articles, adjectives, or demonstratives plus nouns in German (Bürki, Sadat, Dubarry, & Alario, 2016; Heim, Friederici, Schiller, Rüschemeyer, & Amunts, 2009; Schiller & Caramazza, 2003; Schiller & Costa, 2006; Schriefers & Teruel, 2000), Dutch (La Heij, Mak, Sander, & Willeboorsde, 1998; Schiller & Caramazza, 2003, 2006; Starreveld & La Heij, 2004; van Berkum, 1997), and Czech (Bordag & Pechmann, 2008). The GCE was also found when considering referential processing through pronouns in Croatian (Costa, Kovacic, Fedorenko, & Caramazza, 2003). However, the effect has been consistently absent when, (1) bare nouns are used to name the pictures (e.g., "*tafel*"; Finocchiaro et al., 2011; La Heij et al., 1998; Starreveld & La Heij, 2004), or (2) the form of the determiners does not vary across genders (e.g., Schiller & Caramazza, 2003, 2006). The latter is the case of plural definite articles in German and Dutch (in the singular, the articles are "*der*" [masculine], "*die*" [feminine], "*das*" [neuter] in German and "*de*" [common] and "*het*" [neuter] in Dutch, but in the plural the articles are "*die*" [German] and "*de*" [Dutch] for all gender values). It is also the case with diminutive noun forms in Dutch (for which the article "*het*" is used regardless of gender). Thus, the GCE is only obtained when the article varies across gender values (the singular and standard [non-diminutive] conditions). Yet, in line with WEAVER++, it could be said that the GCE is not observed in the plural and diminutive forms due to the absence of different determiner forms across gender values (i.e., there is no agreement to be fulfilled), which makes gender selection unnecessary. However, it has been systematically shown in other studies that gender is indeed being selected in these cases. More specifically, authors looking at this have often used another task, the singular-plural paradigm, to draw conclusions on the effects observed in the PWI paradigm when these were not sufficiently clear to properly interpret certain outcomes.

In the singular-plural paradigm (as well as in its adaptation to the production of diminutives and standard forms), participants have to carry out a simple picture-naming task using noun phrases in the Dutch or German plural (or, in the case of the abovementioned adaptation, in the diminutive Dutch form). Results systematically show faster responses when the article is the same across number values and noun forms (Janssen & Caramazza, 2003; Schriefers, Jescheniak, & Hantsch, 2002, 2005; Spalek & Schriefers, 2005). Regarding number, this means that the plural articles "*die*" (German) and "*de*" (Dutch) entail faster responses when naming feminine (German) and common (Dutch) plural nouns, since both require the same article in the singular. For instance, for the feminine German noun "*Tür*" ("door"), we would say "*die Tür*" in the singular and "*die Türen*" in plural (but for a masculine noun, we would say "*der Tisch*" in the singular vs. "*die Tische*" in the plural). In this sense, a cost is consistently observed in the plural when there is a mismatch with the singular (e.g., the singular masculine in German, "*der*"). The same applies to Dutch noun form, with consistent faster responses for neuter diminutive forms than for common diminutive forms. This because neuter diminutive forms require the same article in their standard form and so, the neuter noun "*kasteel*" (castle) when transformed into the diminutive (marked by suffix "*-je*") maintains the article "*het*" ("*het Kasteel*" and "*het Kasteelje*"), but a common gender noun such as "*stoel*" (chair) sees "*de*" replaced by "*het*" ("*de boek*" vs. "*het boekje*").

As argued by Caramazza, Miozzo, Costa, Schiller, and Alario (2001), this set of results strongly supports the claims that: (1) gender is being selected even in the absence of agreement in the plural and diminutive forms, since singular/plural and standard/diminutive determiner forms compete for selection on the basis of gender, and (2) the locus of the GCE

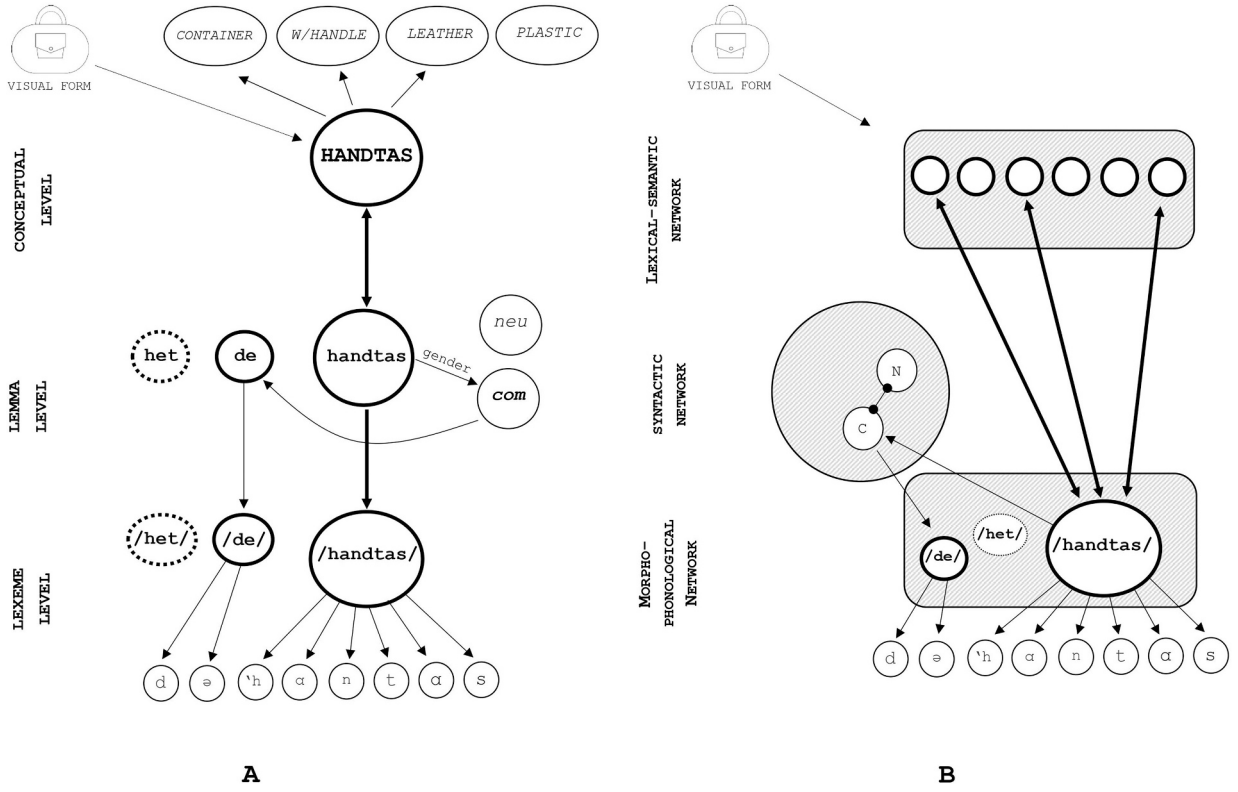

Fig. 1. The processing of gender according to the classical models of language production.
Note. Representation of the Dutch noun phrase "*de handtas*" (the handbag) during language production according to the WEAVER++ (A) and Independent Network (B) models. Continuous lines represent the flow of information as well as the selection of nodes. Discontinuous lines represent unselected nodes. According to WEAVER++, gender selection takes place after the lemma node for "*handtas*" activates the common gender node. The selection of this gender node is necessary for the selection of the correct definite article "*de*". (B) C = Common gender, N = Neuter gender. According to the Independent Network model, abstract gender selection occurs thanks to the flow of activation coming from the lexeme level. Competition between gender values takes place and leads to the inhibition of the node not reaching the threshold for selection. Figures adapted from Levelt et al. (1999) and Caramazza (1997).

is hence at the level of phonological encoding of the determiners. Critically, the authors also state that because the GCE reflects competition between determiners, gender is a feature that is made automatically available prior to lexical selection without any reference to thresholds for selection and without competition occurring between gender nodes as Schriefers (1993) would claim (e.g., Caramazza et al., 2001; Schiller & Caramazza, 2003).

Similar remarks regarding the absence of competitive mechanisms in the selection of certain features are made when analysing in more detail the elements of agreement that show the GCE. Briefly, when participants use agreement elements whose form variation is made through bound morphemes ("*groene*" [common] vs. "*groen*" [neuter]), the GCE is unreliable. Schiller and Caramazza (in Dutch and German, 2003), Costa et al. (in Croatian, 2003), and Schiller and Costa (in German, 2006) failed to obtain any type of gender effect when participants named the pictures using bound morphemes (attached to adjectives, possessive pronouns, and indefinite articles, respectively). However, the GCE was consistently obtained with the same materials when freestanding elements such as definite articles were used (e.g., "*de*" and "*het*"). As in their previous claim regarding the nature of the mechanisms underlying gender selection, Schiller and colleagues bore in mind the ideas of Lapointe and Dell (1989) and argued that bound morphemes are not subject to selection by competition. Gender agreement for gender inflections would thus involve subsequent "phonological transformations" of the stem rather than the selection of independent morphemes by competition-based mechanisms (Lemhöfer, Schriefers, & Jescheniak, 2006). However, there is inconsistent evidence in the field here. Schriefers (1993) and Bordag and Pechmann (2008) both found a clear GCE with Dutch and Czech speakers when using bound morphemes (attached to adjectives and ordinals, respectively), suggesting classical competitive mechanisms of selection. Similarly, there is evidence within the singular-plural paradigm that supports the idea of competition. This simpler naming task consistently shows that bound morphemes are selected in an analogous manner to freestanding forms (Janssen & Caramazza, 2003; Lemhöfer et al., 2006; Schriefers et al., 2002, 2005; Spalek & Schriefers, 2005). The gender value whose plural morphemic form coincides with the singular form shows lower response times (RTs), but gender values in which plural and singular morphemic forms differ show a disadvantage (higher RTs). Then, for instance, in German, adjectives mark gender through bound morphemes, so the translation for the adjective "big" is "*großer*" (masculine), "*große" (feminine)*, and "*großes" (neuter)*. The plural, however, always coincides with the feminine ending "-e" ("*große*") regardless of gender. When participants use the plural for the feminine ("*große*"), faster responses are observed in comparison to the feminine singular, and also in comparison to plural masculine and neuter.

Certainly, the ambiguity of the results within the PWI paradigm stands out when compared to the more consistent results in the singular-plural paradigm. For this reason, it may be that the PWI paradigm, due to its inherent complexity when mixing language comprehension (of the distractor) and production (of the target), is not sensitive enough to capture small lexical access effects. In this sense, competition and priming processes might be more difficult to detect behaviourally by measuring onset naming latencies for elements that occur in non-initial positions of the utterance than for elements that are in initial position (Lemhöfer et al., 2006). Indeed, bound morphemes tend to appear later in the utterance (e.g., compare a freestanding to a bound morpheme: "***der/die***" vs. "*ein* **(-)** */eine*"). Directly related to this, an inspection of the

literature suggests that the higher the number of agreeing elements available as a possible response, the higher the likelihood of observing a GCE with bound morphemes (see Table 1). When only one element is used (e.g., the indefinite article in German "*ein/e*" [M/F]; Schiller & Costa, 2006) the stem is always the same (e.g., "*ein-*"), and this can indeed produce a task-dependent strategy or tendency to initiate the answer as soon as possible by relying on the use of the same stem. This would mask potential differences in processing that might be reflected in the RTs. Indeed, when participants have to choose between two or more stems (e.g., four different colour adjectives in Schriefers, 1993; two different ordinals in Bordag & Pechmann, 2008) the GCE might be easier to observe.[2] In fact, almost all studies on this issue with the singular-plural paradigm have used more than one stem. When Jescheniak, Schriefers, and Lemhöfer (2014) included only one stem (only one adjective) to test this specific hypothesis, they failed to obtain competitive effects. The RTs of those studies with only one stem are in line with this explanation, since they are consistently lower, even when the same materials (pictures) are used (so, participants would indeed be producing faster responses, perhaps reflecting this task-dependent strategy; see Jescheniak et al., 2014).

Finally, when looking at the results of the PWI paradigm with another language family, the Romance branch, the scenario gets slightly more complicated. On the one hand, results with native speakers of Portuguese show that when bare nouns are produced, gender competitive effects can be obtained (Sá-Leite, Haro, Comesaña, & Fraga, 2021). This could have striking repercussions for the way gender representation is conceived: if effects of gender are indeed obtainable with bare nouns, then (1) the GCE is at least partially located at the selection of a gender node, either through facilitative priming (WEAVER++) or competition (IN model); (2) gender is most probably selected without the need for agreement, contrary to the tenets of WEAVER++. However, on the other hand, studies using bare nouns in Spanish and Italian have obtained the effect in an unexpected direction: a gender incongruency effect (GIE; faster RTs for target-distractor pairs of different gender; Cubelli, Lotto, Paolieri, Girelli, & Job, 2005; Paolieri et al., 2010; Paolieri, Lotto, Leoncini, Cubelli, & Job, 2011). This effect is quite surprising, since competition arises when nouns of the same gender are processed together and hence, activation converges on the same gender node. Ultimately, it cannot be interpreted as reflecting processes of competition or priming between gender values, but rather as competition between noun lemmas as a function of gender similarity. Yet, Finocchiaro et al. (2011) failed to replicate the effect in a series of experiments in Italian, Spanish, and French.[3] These mixed results on Romance languages are quite puzzling and make us question once again how truly appropriate the PWI paradigm is to measure certain features of lexical access.

We are thus faced with the following situation: there is a body of work based on the PWI paradigm that aimed to study grammatical gender selection through gender effects but failed to do so in Germanic and Slavic languages, since only determiners seem to be competing. Due to the absence of genuine gender effects, claims with significant repercussions for language processing have been proposed, namely, that any type of competitive or facilitative priming effect underlying the selection of grammatical gender cannot be observed experimentally because gender is likely to be selected automatically. Hence, and contrary to the predictions of both WEAVER++ and the IN model, the GCE would not be located at the level of gender node selection. The scenario with bound morphemes could be thus explained by assuming that they are processed through phonological transformations. The mixed results for Romance languages are also troublesome and might constitute either unfavorable evidence for the claim of automatic gender selection (i.e., the GCE in Portuguese), or controversial evidence supporting different mechanisms that operate gender selection depending on the language family (i.e., mainly the GIE).

We believe that a statistical analysis of the robustness of the effects here discussed is the first necessary step to be taken before stablishing reliable conclusions and proposing alternative ways of disentangling the present results. For this reason, we conducted a meta-analysis to assess the GCE and hence included all available studies testing Germanic and Slavic languages, as well as all those testing Romance languages but limited to bare nouns. We then determined the size of the "GCE" in Germanic and Slavic languages and assessed its robustness depending on the presence of agreement (noun phrases vs. bare nouns) and the type of agreement element included in the naming phrase (freestanding vs. bounded). Regarding Romance languages, we assessed the size of the competitive effects (GIE or GCE) when considering bare nouns. We expected the following results: a significant GCE for Germanic and Slavic languages limited to the presence of agreement but greater for freestanding than bounded elements, as well as a non-significant effect of gender (in any direction) for Romance languages.

Finally, due to the controversial results obtained for Romance languages and the importance of the theoretical claims made here, we assessed a possible publication bias. As a way of reducing this bias, we introduced results with Romance languages from our own lab not yet published.

2. The present study

2.1. Method

2.1.1. Literature search

We conducted a comprehensive and systematic search of two online databases: APA PsycINFO and Web of Science (WoS). The search included the key words: "gender congruency", "grammatical gender", and "picture word interference", the first of these on its own, and the other two combined. We obtained 76 results from APA PsycINFO and 85 results from WoS.[4] From a total of 161 studies, we deleted duplicates with the RefWorks® citation software and obtained 100 different studies of interest. By reviewing the titles, the abstracts, the keywords, plus the full-text of the studies where necessary, we applied the following criteria for inclusion:

a) The study constitutes an empirical exploration of grammatical gender representation and processing during the lexical access of nouns, and entirely avoids the inclusion of nouns with natural or semantic gender in the stimuli list (e.g., Deutsch & Dank, 2018).
b) The study features one or various PWI paradigm tasks without variations on the classical procedure (i.e., fixation cross, optional blank screen, stimuli presentation). Hence, excluded are studies that included extra stimuli as primes (e.g., Alario, Matos, & Segui, 2004) or linguistic instructions preceding the target stimuli (e.g., Finocchiaro & Caramazza, 2006).

[2] There is one study with the PWI task (Schiller & Caramazza, 2003) that included more than one stem and nonetheless failed to obtain significant results.

[3] We are not interested in studies testing agreement contexts in Romance languages. In these (with the exception of Portuguese, see Sá-Leite et al., 2020b), the GCE cannot be explored due to certain particularities of the determiner system of these languages which make the results systematically null (see the late selection hypothesis, Miozzo & Caramazza, 1999; see also Bürki, Besana, Degiorgi, Gilbert, & Alario, 2019).

[4] We excluded the following categories from the WoS database search: *Business, management, audiology speech language pathology, education research, psychology applied, psychology social, social sciences interdisciplinary, film radio television, health care science services, medicine general internal, otorhinolaryngology, public environmental occupational health, radiology nuclear medicine medical imaging*, and *urban studies*.

Table 1
Summary of the studies using bound morphemes.

Study	Experiment	Type of task	Word class	Language	Number of stems	Effect of competition
Bordag and Pechmann (2008)	2 & 3	PWI	Ordinals	Czech	2	Yes
Costa et al. (2003)	2 & 3	PWI	Possessives	Croatian	1	No
Schiller and Caramazza (2003)	1b	PWI	Adjectives	German	2	No
	4a	PWI	Adjectives	Dutch	4	No
Schiller and Costa (2006)	4a	PWI	Indefinite articles	German	1	No
Schriefers (1993)	2	PWI	Adjectives	Dutch	4	Yes
Jescheniak et al. (2014)	Unique	S-P	Adjectives	Dutch	1	No
Lemhöfer et al. (2006)	Unique	S-P	Adjectives	Dutch	2	Yes
Schriefers et al. (2005)	3	S-P	Adjectives	German	1	Yes (partially)

Note. PWI = Picture-Word Interference paradigm; S-P = Singular-Plural paradigm. Effects for the PWI paradigm refer to the gender congruency effect, whereas for the S—P paradigm refer to faster responses for the form shared between the plural and a certain gender in the singular (e.g., the indefinite article in German, "*eine*" [plural, all gender values] – "*eine*" [feminine]) and slower responses for the forms not shared between the plural and the rest of gender values in the singular (e.g., "*eine*" – "*ein*" [masculine]). Importantly, Jescheniak et al. (2014) is an identic replication of Lemhöfer et al. (2006) to test the hypothesis on the number of stems. Schriefers et al. (2005) only used one stem and found faster responses for the shared form (feminine in plural in comparison to singular) but not slower responses for the non-shared forms (masculine and neuter in plural in comparison to singular and in comparison to the feminine plural).

c) Gender congruency was included as one of the experimental conditions.
d) The GCE explored in the study was restricted to the native language or dominant language of the speakers, and the study did not address exclusively the cross-linguistic GCE, i.e., the effect that reflects the cross-linguistic interaction between translations of the same or opposite gender value during lexical access in bilinguals.
e) Participants were neurologically typical and did not present any language impairment (e.g., aphasias).
f) Participants were speakers of Germanic, Romance, or Slavic languages, since Semitic languages such as Hebrew (e.g., Dank & Deutsch, 2015) display a very different set of word-formation rules that may affect the comparability of results with the other language families.
g) Participants were asked to name pictures using either a bare noun, a noun phrase, or pronouns (this excludes any task variations in which participants had to carry out extra cognitive operations such as producing articles without nouns, e.g., Starreveld & La Heij, 2004).
h) Since the focus of our meta-analysis was not exploring the late selection hypothesis (Miozzo & Caramazza, 1999), the studies featured speakers of Germanic and Slavic languages with the naming instructions defined on the previous criterion, or speakers of Romance languages who used bare nouns to name the pictures (i.e., this excludes studies featuring speakers of Romance languages using noun phrases, e.g., Alario & Caramazza, 2002; Miozzo & Caramazza, 1999).
i) The aim of the study was not to explore grammatical gender selection in derived or compound nouns (e.g., Lorenz & Zwitserlood, 2016).

After the application of the inclusion criteria, we identified 12 studies of interest (Bordag & Pechmann, 2008; Bürki et al., 2016; Costa et al., 2003; Cubelli et al., 2005; La Heij et al., 1998; Paolieri et al., 2010, 2011; Schiller & Caramazza, 2003; Schiller & Costa, 2006; Schriefers, 1993; Schriefers & Teruel, 2000; van Berkum, 1997). An inspection of the reference lists of these papers was subsequently conducted and four other studies were deemed eligible (Finocchiaro et al., 2011; Heim, Friederici, et al., 2009; O'Rourke, 2007; Schiller & Caramazza, 2006). Another study was published during this process, which we also included (Sá-Leite et al., 2021). Importantly, when several versions of the same study were found (e.g., journal paper and dissertation: O'Rourke, 2007, 2009), we considered only the journal version. All the authors in the field were contacted by email to request the raw data from their published works as well as unpublished data. Raw data was obtained from Cubelli et al. (2005), the fourth experiment of Finocchiaro et al. (2011), Paolieri et al. (2010, 2011), Sá-Leite et al. (2021), and van Berkum (1997). Unpublished data were included as a new entry (Sá-Leite, Oliveira, Soares, Carreiras, & Comesaña, 2017). In total, 18 studies (17 published papers, one unpublished work with 3 experiments) were carefully analysed, as described in Table 2.

2.1.2. Meta-analytic approach

Our search identified 43 experiments from 18 studies with 93 comparisons of interest for the GCE. From these comparisons, we computed Hedges' *g* as the effect size measure (for computations and formulae, see the Supplemental Materials). In many experiments, the same sample provided more than one effect size. To avoid problems arising from the non-independence of the effect sizes, we used the robust variance estimation method (Hedges, Tipton, & Johnson, 2010), which considers the correlation between observations and thus does not require independent effect sizes (for technical details on the robust variance estimation method, see also the Supplemental Materials). All the confidence intervals reported below are 95% *CI*s. Data entered into the meta-analysis are available at the following link: https://osf.io/myx45/?view_only=e6e3d1d10c5444d196c916a9dffbe8e2.

3. Results

We conducted a series of analyses to test the predictions described in the Introduction. First, we assessed the overall size of the GCE. The effect with all the comparisons of interest was small yet different from zero, $g = 0.112$, $SE = 0.044$, CI [0.023, 0.201], $t(42) = 2.54$, $p = .015$. Heterogeneity between effects was substantial, $I^2 = 81.48\%$, and thus we tried several moderators.

First, we computed the overall effect with Germanic/Slavic languages ($k = 56$). The effect was different from zero, $g = 0.270$, $SE = 0.044$, CI [0.179, 0.362], $t(26) = 6.07$, $p < .001$. For these 56 comparisons, we considered the impact of the agreement context. The GCE was higher when there was agreement, $g = 0.321$, $SE = 0.050$, CI [0.217, 0.425], than when there was no agreement, $g = -0.022$, $SE = 0.063$, CI [−0.154, 0.110], $t(21) = 3.75$, $p = .001$. As expected, the effect was limited to the presence of agreement and hence was only different from zero when there was agreement, $t(21) = 6.44$, $p < .001$. To further explore the conditions that affect the GCE, we selected the comparisons from experiments with Germanic/Slavic languages with an agreement context ($k = 41$) and tested the role of the type of agreement on the GCE. The GCE was similar between freestanding elements, $g = 0.424$, $SE = 0.061$, CI [0.296, 0.553], and freestanding plus bound morphemes, $g = 0.425$, $SE = 0.125$, CI [0.163, 0.688], $t(18) = 0.01$, $p = .993$. These two effects were different from zero, $t(18) = 6.93$, $p < .001$ and $t(18) = 3.40$, $p = .003$, respectively. The effects were higher in these two conditions than in the bound morpheme only condition, $g = 0.042$, $SE = 0.040$, CI [−0.042, 0.125], $t(18) = 5.24$, $p < .001$, and $t(18) = 2.92$, $p = .009$, respectively. Thus, the effect was greater for freestanding elements, yet, unexpectedly, it was almost similar to zero and non-significant for

Table 2
Summary of all reviewed studies on the gender congruency effect in the meta-analysis.

	Exp.	Part. (n)	Language	Language Family	Agreement context	Type of Phrases	SOA	Gender effects
Bordag and Pechmann (2008)	E1	32	Czech	Slavic	Yes	NP (dem + N)	0	GCE
	E2	16	Czech	Slavic	Yes	NP (adj + N)	0	GCE
	E3	14	Czech	Slavic	Yes	NP (adj + N)	0	GCE
Bürki et al. (2016)	E1	18	German	Germanic	Yes	NP (df + N)	0	GCE
Costa et al. (2003)	E1	20	Croatian	Slavic	Yes	Pronoun	0	GCE
	E2	20	Croatian	Slavic	Yes	NP (adj + N)	0	Null
	E3	19	Croatian	Slavic	Yes	NP (adj + N)	225	Null
Cubelli et al. (2005)	E1	28	Italian	Romance	No	Bare noun	0	GIE
	E2	28	Italian	Romance	Yes	NP (df + N)	0	Null
	E3	28	Italian	Romance	No	Bare noun	0	GIE
	E4	28	Italian	Romance	No	Bare noun	0	GIE
Finocchiaro et al. (2011)	E1	24	Italian	Romance	No	Bare noun	0	Null
	E2	28	Italian	Romance	No	Bare noun	0	Null
	E3	30	Spanish	Romance	No	Bare noun	0	Null
	E4	24	French	Romance	No	Bare noun	0	Null
	E5	20	German	Germanic	No	Bare noun	0	Null
	E6	20	Dutch	Germanic	No	Bare noun	0	Null
Heim, Friederici, et al. (2009)	E	14	German	Germanic	Yes	NP (df + N)	0	GCE
La Heij et al. (1998)	E1	20	Dutch	Germanic	No/Yes	Bare noun/NP (df + N)	0	Null/GCE
	E2	20	Dutch	Germanic	No/Yes	Bare noun/NP(df + N)	0	GCE*/GCE
	E3b	16	Dutch	Germanic	No/Yes	Bare noun/NP (df + N)	0	Null/GCE*
O'Rourke (2007)	E1	16	Spanish	Romance	No/Yes	Bare noun/ NP (df + N)	0	Null/Null
Paolieri et al. (2010)	E1	16	Italian	Romance	No	Bare noun	0	GIE
	E2	20	Spanish	Romance	No	Bare noun	0	GIE
Paolieri et al. (2011)	E1	36	Italian	Romance	No	Bare noun	0	GIE
	E2	36	Italian	Romance	No	Bare noun	0	GIE
Sá-Leite et al. (2017)	E1	48	Portuguese	Romance	No	Bare noun	0	Null
	E2	48	Portuguese	Romance	No	Bare noun	0	Null
	E3	80	Portuguese	Romance	No	Bare noun	0	Null
Sá-Leite et al. (2021)	E1-cond1	36	Portuguese	Romance	No	Bare noun	0	GCE
	E2	48	Portuguese	Romance	No	Bare noun	0	Null
Schiller and Caramazza (2003)	E1a	27	German	Germanic	Yes/No	NP (df + N) – S/Pl	0	GCE/Null
	E1b	25	German	Germanic	Yes/No	NP (adj + N) – S/Pl	0	Null/Null
	E1c	26	German	Germanic	Yes/No	NP (df + N + adj) – S/Pl	0	GCE/Null
	E2a	17	Dutch	Germanic	Yes/No	NP (df + N) – S/Pl	-100, 0, +100	GCE/GCE
	E2b	18	Dutch	Germanic	Yes/No	NP (df + N) – S/Pl	0	GCE/Null
	E3	26	Dutch	Germanic	Yes/No	NP (df + N) – S/Pl	0	GCE/Null
	E4a	8	Dutch	Germanic	Yes	NP (adj + N)	-100, 0	Null
	E4b	15	Dutch	Germanic	Yes	NP(df + N + adj)	0	GCE*
Schiller and Caramazza (2006)	E1	28	Dutch	Germanic	Yes	Noun phrase (df + N)	0	GCE
	E2	19	Dutch	Germanic	Yes	Noun phrase (df + N)	0	GCE
Schiller and Costa (2006)	E1a	20	German	Germanic	No	Noun phrase (in + N)	0	Null
	E1b	20	German	Germanic	No	Noun phrase (df + N)	0	GCE

(*continued on next page*)

Table 2 (*continued*)

	Exp.	Part. (n)	Language	Language Family	Agreement context	Type of Phrases	SOA	Gender effects
Schriefers and Teruel (2000)	E1	16	German	Germanic	Yes	Noun phrase (df + adj + N)	-150, 0, +150, +300	GCE +150
	E2	16	German	Germanic	Yes	Noun phrase (df + N)	-225, -150, -75, 0, +75, +150, +225, +300	GCE +75
Schriefers (1993)	E1	18	Dutch	Germanic	Yes	Noun phrase (df + adj + N)	-200, 0, +450	GCE -200, 0
	E2	18	Dutch	Germanic	Yes	Noun phrase (adj + N)	-200, 0, +450	GCE 0
van Berkum (1997)	E2	48	Dutch	Germanic	Yes	Noun phrase (df + N)	0	GCE

Note. Exp. = Experiment. Part. = Participants. Agreement context: Yes = presence of agreement; No = absence of agreement. Type of Phrases: NP = noun phrase; df = definite article; in = indefinite article; adj = adjective; dem = demonstrative; pron = pronoun; N = Noun; S = Singular; Pl = Plural. SOA = Stimuli Onset Asynchrony. Gender effects: GCE = Gender Congruency Effect; GIE = Gender Incongruency Effect; Null = Absence of effects. * = significant only by participants or items. Bürki et al. (2016) and O'Rourke (2007) were not included in the analyses due to an insufficient amount of reported statistical information and the lack of raw data.

bound morphemes.

Then, we computed the overall effect with Romance languages (k = 37). There was a negative GCE (or GIE effect) which was not canceled by the null results and the GCE, $g = -0.108$, $SE = 0.049$, CI [−0.212, −0.003], meaning that participants were slower responding to gender congruent items. Although the effect was not especially robust, it was different from zero, $t(14.67) = 2.20$, $p = .044$.

We also conducted a publication bias analysis to assess the possible overestimation of the effects. Fig. 2 shows a funnel plot with all the effect sizes. A funnel plot is a scatter plot with the effect sizes represented in the X axis and their SE in the Y axis. As the precision of the effects decreases (i.e., the SE gets higher at the bottom part of the Y axis) it is expected that the effects would scatter widely along the X axis. As the SE increases, a narrower spread of the effects around the true effect size is expected. Thus, a "funnel" scatter of the data would show no bias. Alternatively, if the bottom left part of the plot seems empty, that would indicate that studies with low precision (likely due to low n) and effects lower than usual in the area were not included in the meta-analyses, probably because they were not published. This distribution of the effects can be measured with Egger's regression test for funnel plot asymmetry. When all the comparisons of interest were included, a regression showed asymmetry and risk of bias, $z = 6.55$, $p < .001$.

Since we found large differences in the GCE between experiments with different languages, we also conducted separate analyses per language family. For Germanic/Slavic languages the asymmetry of the funnel plot was significant, $z = 2.33$, $p = .020$. However, assessing the effect in this way is not optimal, since it mixes two types of naming instruction (noun phrases and bare nouns) and hence blurs the effect of determiner congruency (the one which interests us, and which differs from a genuine GCE explored with bare nouns). When considering only experiments in which an agreement fulfilment was required, the asymmetry disappeared: $z = 1.08$, $p = .280$ (see Fig. 3, Panel A). On the other hand, with Romance languages and bare nouns, the asymmetry is also absent, $z = 0.59$, $p = .553$ (see Fig. 3, Panel B). This means that it is unlikely that the effects computed here overestimated the actual effects.

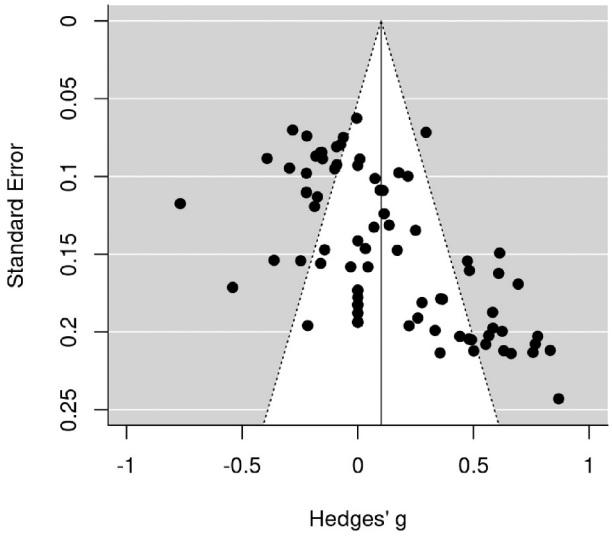

Fig. 2. Funnel plot with all the effects.

4. Discussion

In this study, we addressed the state of the art of grammatical gender retrieval during native language production by focusing principally on the so-called GCE obtained with the PWI paradigm. The GCE has been described as a determiner congruency effect for Germanic and Slavic languages, although studies show inconclusive results when it comes to the type of agreement involved in the emergence of the effect. Conversely, in certain studies exploring Romance languages, either a GCE emerges with bare nouns, indicating a genuine effect at the level of gender node selection, or the effect emerges in the opposite direction, as a GIE. Still, null effects have also been obtained by other studies in the same Romance languages. Importantly, the complexity of the PWI paradigm has been pointed out as a possible obstacle to accurately detect small effects of competition or facilitative priming, especially when it comes to gender. Through a meta-analysis, we intended to assess the robustness of the effects of either determiner forms or gender values in Germanic/Slavic and Romance languages, as a way to understand: 1) the nature of the mechanisms underlying gender selection, and 2) the nature of the mechanisms underlying bound morpheme selection in Germanic languages.

The results showed that the overall GCE for Germanic and Slavic languages is small but highly significant ($g = 0.270$). When considering the presence of agreement, we were able to confirm that the GCE consistently emerges when elements of agreement that vary across gender values are produced, but not when bare nouns or elements of agreement that do not vary across gender values are used (e.g., definite articles in their plural form in the case of German and Dutch). Importantly, the meta-analysis shows robust evidence for the effect when freestanding elements are present, either alone with the noun or with other words displaying bound morphemes (e.g., definite articles plus adjectives plus nouns). Yet, the effect is not significant for noun phrases with only bound morphemes plus nouns. Overall, the results with these languages support the idea that the GCE is a determiner congruency effect of moderate size ($g = 0.425$). Presumably, they are likewise in line with the claim that gender is selected automatically, and that this selection does not bear competitive or facilitative priming effects based on the degree of activation reached by gender values (Caramazza et al.,

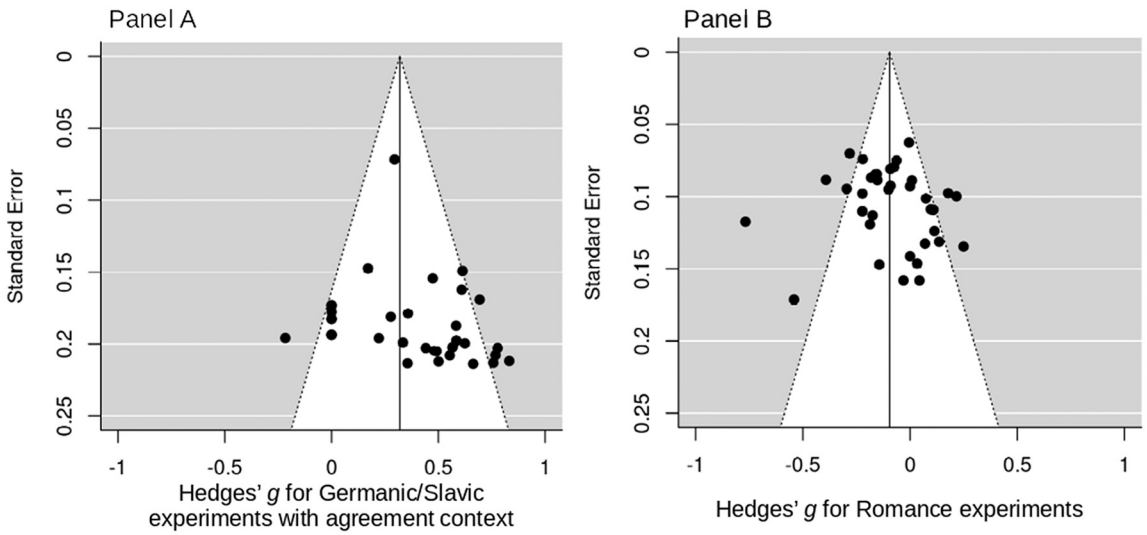

Fig. 3. Funnel plots for experiments with Germanic/Slavic Languages (Panel A) and with Romance Languages (Panel B).

2001). The results also revealed a GIE when speakers of Romance languages name the pictures using bare nouns, but it is small and barely significant ($p = .044$). Note that this effect defies the current models of lexical access. Finally, we did not detect any publication bias, either in studies of the "determiner" congruency effect with Germanic and Slavic languages, or in studies of gender effects with Romance languages.

These findings allow us to paint the following picture: the PWI paradigm does not reflect effects of gender in Germanic and Slavic languages, whereas in Romance languages a puzzling gender incongruency effect has been found, challenging the current views on the architecture and functioning of lexical access. However, there are important considerations to be made here.

On the one hand, regarding the results with Germanic and Slavic languages, although competition between elements of agreement is the only type of effect that can be identified through the PWI paradigm, we are unsure about how this can be taken as evidence for the absence of gender effects and the automatic nature of gender selection. In other words, competition between determiners cannot be taken as evidence for the absence of competition or priming between gender features during gender selection. This is especially relevant when restricting ourselves to only one paradigm, which is fairly complex and might be inadequate to detect effects that are very small. Note that our meta-analysis reports a determiner congruency effect that is within the moderate range of effects, with a *g* of 0.425. Indeed, gender-based effects are very small (bear in mind that the GIE for Romance languages, in absolute value, showed a *g* of only 0.108 and is quite heterogeneous) and hence, the complexity of the PWI paradigm may certainly be a problem to detect them. In this sense, the GCE obtained with bilinguals entails faster responses for translations that have the same gender in comparison to those that have different gender in simple naming and translation tasks. This genuine cross-linguistic effect of gender is observed consistently even for language pairs belonging exclusively to the Germanic family (e.g., German and Dutch, Lemhöfer, Spalek, & Schriefers, 2008), yet it is smaller than the determiner congruency effect (it shows a mean *g* of 0.24, see Sá-Leite, Luna, Fraga, & Comesaña, 2020). The use of other methodologies that would allow to better capture small effects such as those related to gender is thus recommended within this field. In this sense, we might mention here the work of Heim, Friederici, et al. (2009), who conducted an fMRI study with the PWI paradigm and native speakers of German using the same materials as in Schiller and Caramazza (2003). Participants were asked to name the pictures using noun phrases (definite article plus noun). They found a standard GCE on the RTs, but critically, they also found that the haemodynamic response (the activation) in the left BA 44 increased faster when the target and the distractor had the same gender. Neuroimaging studies suggest that there is a segregation of the left BA 44 in which the retrieval of phonological forms has been shown to involve its superior portion (e.g., Heim, Alter, & Friederici, 2005; Heim, Opitz, Müller, & Friederici, 2003; Longoni, Grande, Hendrich, Kastrau, & Huber, 2005; Marangolo, Piras, Galati, & Burani, 2006; Miceli et al., 2002), while syntactic processing of features such as grammatical gender or word class seem to emerge in its inferior portion (e.g., Heim et al., 2003, 2005; Longoni et al., 2005). The haemodynamic effects that Heim, Friederici, et al. obtained were restricted to the inferior portion of the left BA 44, and did not emerge in the superior portion, in which phonology and bound morphemes are encoded. The authors' interpretation is that this effect may reflect gender selection that occurs prior to the selection of the determiner form. In the same year, Heim, Eickhoff, Friederici, and Amunts (2009) conducted a German naming task with priming, in which prime and target varied as a function of semantic relation, gender congruency, and phonological overlap. Critically, participants were asked to name pictures using bare nouns. Although the results concerning RTs were not sufficiently clear, the fMRI revealed that the same area of the BA 44 showed modulations regarding the gender congruency of prime and targets. These differed significantly from the effects of phonological overlap, and were similar to the ones reported in Heim, Friederici, et al. (2009). This is in line with the existence of effects linked to the selection of abstract gender features even in a Germanic language. In sum, although our meta-analysis supports the existence of a determiner congruency effect in the studies exploring Germanic and Slavic languages, we would refrain from excluding that gender is selected at the lemma level. Absence of evidence cannot be interpreted as evidence for the absence, and there is in fact evidence against such affirmation.

Moreover, the determiner congruency effect was only significant when freestanding elements were present. These results match the ideas of Lapointe and Dell (1989), later supported by Schiller and Caramazza (2003), according to whom bound morphemes are not selected through competitive mechanisms like freestanding elements are, but instead through phonological transformations on the stem of the agreeing word. In any case, again, we would refrain from making clear-cut affirmations here, as the number of studies on the bound morphemes issue is rather small. Thus, this conclusion must be taken with caution, especially considering once more criticisms of the complexity of the PWI paradigm and the possible repercussions that such complexity can have when seeking to capture small effects.

It is precisely because of certain inadequacies in the PWI paradigm that many authors have heavily relied on the singular-plural paradigm to: (1) determine whether gender is being selected when plural and

diminutive nouns are produced, and (2) assess the mechanisms underlying the selection of bound morphemes. It would be interesting, thus, to conduct a meta-analysis of the robustness of the effects obtained in this paradigm, and even compare the sizes of these effects with those of the PWI paradigm. However, there are not enough studies in the literature to properly assess the results, to obtain effects that are different from zero, and hence to make a proper interpretation of the robustness and significance of these results (only 10 comparisons of interest are available).[5] Therefore, we are not able to confirm that gender is indeed being selected when an agreement context is absent (plural/diminutive forms), or that there is a competitive effect between bound morphemes of different gender in this paradigm. Critically, the results with this task were the main evidence for Schiller and Caramazza (2003) in their claim that the GCE is a determiner congruency effect. The authors repeatedly found null effects when using plural definite articles whose only form is "*die*" in German and "*de*" in Dutch. We could interpret these results through the tenets of the WEAVER++ model and say that perhaps gender was not being selected because agreement was not necessary, and thus the GCE did not emerge. However, through the singular-plural paradigm the authors confirmed that faster or slower responses depended on gender, suggesting that gender was indeed being selected, even in the plural. Thus, the GCE was not a genuine gender effect dependable on the presence of an agreeing context, but rather an effect of competition between determiners. Given that there are insufficient studies on the singular-plural paradigm to establish a significant effect between freestanding determiners, this interpretation of the GCE lacks the degree of robustness that would be desirable.

Regarding the results with studies exploring Romance languages, we shall note that the GIE is an effect whose observation comes from only one laboratory, which tested native speakers of Italian and Spanish (Cubelli et al., 2005; Paolieri et al., 2010, 2011). Two other laboratories have tried to replicate it in Spanish, Italian, and Portuguese, but failed to do it. A GIE implies that nouns of the same gender compete, rather than facilitate the selection of the same gender value. This is hard to interpret theoretically given the current evidence on gender processing, which suggests the existence of an advantage coming from lexical entries of the same gender in comparison to those of different gender (see Bates, Devescovi, Hernandez, & Pizzamiglio, 1996; Bender, Beller, & Klauer, 2011; Heim, Eickhoff, et al., 2009; Paolieri, Padilla, Koreneva, Morales, & Macizo, 2019; Sá-Leite, Fraga, & Comesaña, 2019; Sá-Leite, Luna, et al., 2020; Vigliocco, Lauer, Damian, & Levelt, 2002). However, our meta-analysis reported a weak GIE. Note that for Romance languages we have a total of 37 effects. Of these, 24 are either zero or extremely small on the positive or negative side, even the significant GCE from Sá-Leite et al. (2021). The greatest effect among these is a g of 0.250 from Finocchiaro et al. (2011), and it is in fact not significant. Yet, the studies that found a significant GIE, those of Cubelli et al. (2005) and Paolieri et al. (2010, 2011), represent the 13 remaining effects. Among these 13, the effect size ranges from $g = -0.091$ to $g = -0.768$. Importantly, the study of Cubelli et al. (2005) found effects ranging between $g = -0.091$ and $g = -0.222$. Ultimately, the meta-analysis reports a significant GIE due to 7 of these 37 effect sizes, this is, those belonging to Paolieri et al. (2010, 2011) studies. That is why even though the studies of Paolieri represent 19% of the effects and the GIE has not been replicated, it was significant in our meta-analysis. It is thus worth mentioning that this kind of erratic pattern of results (non-replicability of the effects resulting in null effects, plus effects in the opposite direction of what was expected; plus effects that have quite a large size when compared to other studies that obtain null results) has been described by many statisticians as a symptom of lack of power due to small participant samples (Type S error and overestimation; see Vasishth & Gelman, 2021). Certainly, lack of power is a cause for the finding of false effects (Brysbaert, 2019). Indeed, the GIE approaches significance especially due to Paolieri et al. (2010) study whose effect sizes range between - 0.39 and - 0.77.[6] That is exactly the study with the smallest samples (two experiments with 16 participants and one with 20). Small samples with designs that have more than one factor when exploring gender effects (which are known to be small) is probably not a good recipe for such a complex paradigm. In short, our meta-analysis offers little support to the GIE and we believe that further replication would confirm this. In this sense, note that the significant effect found by Sá-Leite et al. (2021) also calls for replication. It was an effect of congruency but stands alone in our sample. Therefore, it might either be a fluke or constitute evidence for a different mechanism of gender selection for European Portuguese, which seems unlikely.

All in all, the nature of claims made in the literature on the GCE with Germanic and Slavic languages is highly relevant for an understanding of language production. Such claims involve the mechanisms underlying the selection of grammatical features and morphemes and propose that some of these do not entail any type of priming or competitive processes. Most models of language processing take lexical competition as the basis of language processing (e.g., Becker, 1980; Dell, 1986; Levelt et al., 1999; Vigliocco & Hartsuiker, 2002). Neuroscientific evidence and computational models of language processing also support this conception (see Dijkstra et al., 2019). It is thus surprising that features may exist that are selected "automatically" without yielding any type of priming or competition between nodes. Indeed, there is evidence relying on other type of measurements that supports the idea that gender nodes are actually selected and yielding effects at the level of lemma (Heim, Eickhoff, et al., 2009; Heim, Friederici, et al., 2009). Yet, behaviourally, only the selection of the determiner seems to be reflected in the RTs. Studies on the singular-plural paradigm might provide a window to test claims of these types, especially when concerning bound morphemes. However, there is a notable scarcity of work with this paradigm. Scarcity is also a problem for the current state of the art with Romance languages, which calls urgently for further replication of the controversial GIE to verify what our meta-analysis suggests (i.e., that it is not reliable at all).

We hence recommend the next approaches to disentangle the mixed results observed within the PWI paradigm and unveil the mechanisms underlying gender selection. First, we believe that researchers from different laboratories should conduct further parametric replications (Regula, 1971)[7] with greater sample sizes of (i) the studies of Cubelli et al. (2005), Paolieri et al. (2010, 2011), and Sá-Leite et al. (2021) to determine what is actually happening during gender selection in Romance languages, and (ii) the studies on bound morpheme selection, not only with the singular-plural paradigm, but also with the PWI paradigm. Besides, we believe that further replication of studies using other type of measurement techniques such as event-related potentials

[5] We conducted a meta-analysis with the six studies that we found in the literature using APA PsycINFO (i.e., Janssen & Caramazza, 2003; Jescheniak et al., 2014; Lemhöfer et al., 2006; Schriefers et al., 2002, 2005; Spalek & Schriefers, 2005). We obtained 10 comparisons of interest for the effect with the coinciding gender in plural (faster responses for feminine plural in German, common plural in Dutch), the effect that is mostly reported as significant (in comparison to the cost effect for the other gender values). The results showed an overall significant effect, $g = 0.303$, $SE = 0.083$, CI [0.111, 0.495], $t(7.74) = 3.66$, $p = .007$. Heterogeneity was also substantial, $I^2 = 69.98\%$. We examined the effect of the type of agreement to understand the extent to which freestanding elements indeed showed the plural advantage (and hence, gender was being retrieved) and to determine what was happening with bound morphemes. We found no differences between conditions, freestanding $g = 0.312$, $SE = 0.144$, CI [−0.149, 0.773], bound morpheme $g = 0.243$, $SE = 0.097$, CI [−0.183, 0.669], and both $g = 0.366$, $SE = 0.276$, CI [−3.143, 3.875]. Due to the low number of entries in the meta-analysis, none of the effects were different from zero (see the Supplemental Materials for inferential tests).

[6] If we remove Paolieri et al. (2010) from the sample the GIE is not significant anymore ($k = 34$; $g = -0.059$, $SE = 0.036$, CI [−0.138, 0.020] $t(12.27) = -1.63$, $p = .129$).

[7] We would like to thank Professor Herbert Schriefers for this suggestion.

(Bürki et al., 2016) or fMRI (Heim, Friederici, et al., 2009) is also necessary, as they can better show in such a complex paradigm: (i) the time course of gender selection within lexical access, and (ii) the specific hemodynamic response related to the effects, allowing us to distinguish between the determiner congruency obtained in RTs and a possible smaller gender effect masked within the complexity of the PWI paradigm. The inherent complexity of the PWI paradigm is one of the main issues in this subject and this concern is not novel to other areas. This task has been pointed out as cause of outcome disparity in studies on the word-class and semantic effects. Researchers concluded that, in this paradigm, word-class effects are confounded with greater sized effects based on imageability (nouns are more imageable than verbs; Iwasaki et al., 2008; Mahon et al., 2007). Something similar happens with semantic-based effects, which sometimes arise as interference and other times as facilitation. After a decade of disputes, both facilitation and interference effects seem to be significant and existent but dependent on many variables (type of semantic relation, stimulus onset asynchrony, and visibility). More recent studies with electrophysiological techniques are in line with the idea that both effects exist but become slippery and hard to explore with the PWI paradigm (see Abdel Rahman & Aristei, 2010; Mädebach & Hantsch, 2013; Python, Fargier, & Laganaro, 2018). Therefore, we do believe that the paradigm might be even less suitable to measure small effects of competition, such as those arising from gender (e.g., the cross-linguistic GCE; Sá-Leite, Luna, et al., 2020). For this reason, we highly recommend complementing the paradigm with additional measuring techniques, as said. Besides, converging evidence from alternative paradigms can also be interesting. For instance, we could use the *blocked cyclic naming paradigm* (Kroll & Stewart, 1994), in which cycles of usually four stimuli sharing or not a specific characteristic are presented one after another. The first cycle of stimuli could include four pictures represented by target nouns of the same gender (e.g., masculine gender), followed by a cycle mixing stimuli of different genders (e.g., masculine and feminine gender). Depending on the number of homogeneous cycles that have been presented first before a mismatching stimulus appears, RTs vary within each cycle (previous evidence seems clear-cut for semantic effects, see Python et al., 2018). This could allow us to study gender retrieval as a function of agreement contexts, thus avoiding the presence of distractors and the need to match the time-course of target and distractor. Repeated but controlled presentation and activation of stimuli of the same gender can also make gender effects more visible behaviourally and electrophysiologically.

5. Conclusions

To summarize, the state of the art allows us to affirm that the GCE found for Germanic and Slavic languages is a moderate and significant determiner congruency effect for freestanding elements. Still, we believe that authors should be cautious when interpreting the absence of gender effects in this paradigm as evidence affirming that gender values are not selected. The results with Romance languages show a barely significant GIE at least for Italian and Spanish, although this GIE is dubious and hard to interpret theoretically. Parametric replication is desirable towards a better understanding of how gender is selected during the production of nouns depending on the language family. Further studies on the singular-plural paradigm might help us to understand the exact capacity of the PWI paradigm to detect certain effects, and to reveal which mechanisms underlie bound morpheme selection. Likewise, the use of other type of techniques, such as fMRI, can perhaps help to disentangle the facilitative or competitive effects that may occur during the PWI paradigm, even when these are not reflected in RTs.

Funding

This work was supported by the Government of Spain, Spanish Ministry of Education and Vocational Training, through the Training program for Academic Staff (Ayudas para la Formación del Profesorado Universitario, FPU [FPU16/06983]); the Spanish Ministry of Science and Innovation [research project PID2019-110583GB-I00]; the Galician Government [grant for research groups ED431B 2019/2020]; and the Foundation for Science and Technology of Portugal [IF / 00784/2013 / CP1158 / CT0013]. Finally, the study has also been partially supported by the FCT and the Portuguese Ministry of Science, Technology and Higher Education through national funds and co-financed by FEDER-European Regional Development Fund through COMPETE2020 under the PT2020 Partnership Agreement [POCI-01-0145- FEDER-007653].

CRediT authorship contribution statement

Ana Rita Sá-Leite: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Karlos Luna:** Software, Validation, Formal analysis, Data curation, Writing – review & editing. **Ângela Tomaz:** Validation, Writing – review & editing. **Isabel Fraga:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Montserrat Comesaña:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

We have no conflicts of interest to disclose.

Acknowledgements

We would like to thank Professor F.-Xavier Alario, Professor Kristin Lemhöfer, Professor Daniela Paolieri, and Professor Jos J. A. van Berkum for sharing their data. We would also like to thank Professor F.-Xavier Alario once again for a careful review of this manuscript as well as two anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105060.

References*

Abdel Rahman, R., & Aristei, S. (2010). Now you see it... and now again: Semantic interference reflects lexical competition in speech production with and without articulation. *Psychonomic Bulletin & Review, 17*(5), 657–661. https://doi.org/10.3758/PBR.17.5.657

Alario, F.-X., & Caramazza, A. (2002). The production of determiners: Evidence from French. *Cognition, 82*(3), 179–223. https://doi.org/10.1016/S0010-0277(01)00158-5

Alario, F.-X., Matos, R. E., & Segui, J. (2004). Gender congruency effects in picture naming. *Acta Psychologica, 117*(2), 185–204. https://doi.org/10.1016/j.actpsy.2004.06.003

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. *Perception & Psychophysics, 58*(7), 992–1004. https://doi.org/10.3758/BF03206827

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition, 8*(6), 493–512. https://doi.org/10.3758/bf03213769

Bender, A., Beller, S., & Klauer, K. C. (2011). Grammatical gender in German: A case for linguistic relativity? *Quarterly Journal of Experimental Psychology, 64*(9), 1821–1835. https://doi.org/10.1080/17470218.2011.582128

*Bordag, D., & Pechmann, T. (2008). Grammatical gender in speech production: Evidence from Czech. *Journal of Psycholinguistic Research, 37*(2), 69–85. https://doi.org/10.1007/s10936-007-9060-0

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16, 1–38 https://doi.org/10.5334/joc.72.

Bürki, A., Besana, T., Degiorgi, G., Gilbert, R., & Alario, F.-X. (2019). Representation and selection of determiners with phonological variants. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(7), 1287–1315. https://doi.org/10.1037/xlm0000643

* References marked with an asterisk indicate studies included in the meta-analysis.

Bürki, A., Sadat, J., Dubarry, A.-S., & Alario, F.-X. (2016). Sequential processing during noun phrase production. *Cognition, 146*, 90–99. https://doi.org/10.1016/j.cognition.2015.09.002

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*(1), 177–208. https://doi.org/10.1080/026432997381664

Caramazza, A., Miozzo, M., Costa, A., Schiller, N., & Alario, F.-X. (2001). A cross-linguistic investigation of determiner production. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler* (pp. 209–226). MIT Press.

*Costa, A., Kovacic, D., Fedorenko, E., & Caramazza, A. (2003). The gender congruency effect and the selection of freestanding and bound morphemes: Evidence from Croatian. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1270–1282. https://doi.org/10.1037/0278-7393.29.6.1270

*Cubelli, R., Lotto, L., Paolieri, D., Girelli, M., & Job, R. (2005). Grammatical gender is selected in bare noun production: Evidence from the picture–word interference paradigm. *Journal of Memory and Language, 53*(1), 42–59. https://doi.org/10.1016/j.jml.2005.02.007

Dank, M., & Deutsch, A. (2015). Morphological structure governs the process of accessing grammatical gender in the course of production. *The Mental Lexicon, 10*(2), 186–220. https://doi.org/10.1075/ml.10.2.02dan

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Deutsch, A., & Dank, M. (2018). Morphological structure mediates the notional meaning of gender marking: Evidence from the gender-congruency effect in Hebrew speech production. *Quarterly Journal of Experimental Psychology, 72*(3), 389–402. https://doi.org/10.1177/1747021818757942

Dijkstra, T., Wahl, A., Buytenhuijs, F., van Halem, N., Al-Jibouri, Z., de Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition, 22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

*Finocchiaro, C., Alario, F.-X., Schiller, N. O., Costa, A., Miozzo, M., & Caramazza, A. (2011). Gender congruency goes Europe: A cross-linguistic study of the gender congruency effect in romance and Germanic languages. *Rivista di Linguistica, 23*(2), 161–198.

Finocchiaro, C., & Caramazza, A. (2006). The production of pronominal clitics: Implications for theories of lexical access. *Language & Cognitive Processes, 21*(1–3), 141–180. https://doi.org/10.1080/01690960400001887

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. Erratum in 1(2), 164–165 https://doi.org/10.1002/jrsm.5.

Heim, S., Alter, K., & Friederici, A. D. (2005). A dual-route account for access to grammatical gender: Evidence from functional MRI. *Anatomy and Embryology, 210*(5–6), 473–483. https://doi.org/10.1007/s00429-005-0032-6

Heim, S., Eickhoff, S. B., Friederici, A. D., & Amunts, K. (2009). Left cytoarchitectonic area 44 supports selection in the mental lexicon during language production. *Brain Structure and Function, 213*(4–5), 441–456. https://doi.org/10.1007/s00429-009-0213-9

Heim, S., Friederici, A. D., Schiller, N. O., Rüschemeyer, S. A., & Amunts, K. (2009). The determiner congruency effect in language production investigated with functional MRI. *Human Brain Mapping, 30*(3), 928–940. https://doi.org/10.1002/hbm.20556

Heim, S., Opitz, B., Müller, K., & Friederici, A. D. (2003). Phonological processing during language production: fMRI evidence for a shared production-comprehension network. *Brain Research. Cognitive Brain Research, 16*(2), 285–296. https://doi.org/10.1016/s0926-6410(02)00284-7

Iwasaki, N., Vinson, D. P., Vigliocco, G., Watanabe, M., & Arciuli, J. (2008). Naming action in Japanese: Effects of semantic similarity and grammatical class. *Language and Cognitive Processes, 23*(6), 889–930. https://doi.org/10.1080/01690960801916196

Janssen, N., & Caramazza, A. (2003). The selection of closed-class words in noun phrase production: The case of Dutch determiners. *Journal of Memory and Language, 48*(3), 635–652. https://doi.org/10.1016/s0749-596x(02)00531-4

Jescheniak, J. D., Schriefers, H., & Lemhöfer, K. (2014). Selection of freestanding and bound gender-marking morphemes in speech production: A review. *Language, Cognition and Neuroscience, 29*(6), 684–694. https://doi.org/10.1080/01690965.2012.654645

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*(2), 149e174. https://doi.org/10.1006/jmla.1994.1008

*La Heij, W., Mak, P., Sander, J., & Willeboorsde, E. (1998). The gender-congruency effect in picture word task. *Psychological Research, 61*(3), 209–219. https://doi.org/10.1007/s004260050026

Lapointe, S. G., & Dell, G. S. (1989). A synthesis of some recent work in sentence production. In G. N. Carlson, & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 107–156). Kluwer Academic Publishers.

Lemhöfer, K., Schriefers, H., & Jescheniak, J. D. (2006). The processing of free and bound gender-marked morphemes in speech production: Evidence from Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 437–442. https://doi.org/10.1037/0278-7393.32.2.437

Lemhöfer, K., Spalek, K., & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language, 59*(3), 312–330. https://doi.org/10.1016/j.jml.2008.06.005

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Longoni, F., Grande, M., Hendrich, V., Kastrau, F., & Huber, W. (2005). An fMRI study on conceptual, grammatical, and morpho-phonological processing. *Brain and Cognition, 57*(2), 131–134. https://doi.org/10.1016/j.bandc.2004.08.032

Lorenz, A., & Zwitserlood, P. (2016). Semantically transparent and opaque compounds in German noun-phrase production: Evidence for morphemes in speaking. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01943

Mädebach, A., & Hantsch, A. (2013). Explaining semantic facilitation and interference effects in the picture–word interference task—A rejoinder to Navarrete and Mahon (2013). *Language & Cognitive Processes, 28*(5), 717–722. https://doi.org/10.1080/01690965.2013.770891

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 503–535. https://doi.org/10.1037/0278-7393.33.3.503

Marangolo, P., Piras, F., Galati, G., & Burani, C. (2006). Functional anatomy of derivational morphology. *Cortex, 42*(8), 1093–1106. https://doi.org/10.1016/s0010-9452(08)70221-1

Miceli, G., Turriziani, P., Caltagirone, C., Capasso, R., Tomaiuolo, F., & Caramazza, A. (2002). The neural correlates of grammatical gender: An fMRI investigation. *Journal of Cognitive Neuroscience, 14*(4), 618–628. https://doi.org/10.1162/08989290260045855

Miozzo, M., & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 907–922. https://doi.org/10.1016/S0749-596X(02)00531-4

O'Rourke, P. L. (2007). The gender congruency effect in bare noun production in Spanish. In , *Vol. 15. Coyote papers: Working papers in linguistics* (pp. 66–89). Linguistic Theory at the University of Arizona.

O'Rourke, P. L. (2009). *The nature of syntactic gender processing in spanish: An ERP study*. Doctoral dissertation. University of Arizona. Available from APA PsycInfo. (622059047; 2009-99090-030). Retrieved from https://search.proquest.com/docview/622059047?accountid=17253.

*Paolieri, D., Lotto, L., Leoncini, D., Cubelli, R., & Job, R. (2011). Differential effects of grammatical gender and gender inflection in bare noun production. *The British Psychological Society, 102*(1), 19–36. https://doi.org/10.1342/000712610X496536

*Paolieri, D., Lotto, L., Morales, L., Bajo, T., Cubelli, R., & Job, R. (2010). Grammatical gender processing in romance languages: Evidence from bare noun production in Italian and Spanish. *European Journal of Cognitive Psychology, 22*(3), 335–347. https://doi.org/10.1080/09541440902916803

Paolieri, D., Padilla, F., Koreneva, O., Morales, L., & Macizo, P. (2019). Gender congruency effects in Russian–Spanish and Italian–Spanish bilinguals: The role of language proximity and concreteness of words. *Bilingualism: Language and Cognition, 22*, 112–129. https://doi.org/10.1017/s1366728917000591

Python, G., Fargier, R., & Laganaro, M. (2018). ERP evidence of distinct processes underlying semantic facilitation and interference in word production. *Cortex, 99*, 1–12. https://doi.org/10.1016/j.cortex.2017.09.008

Regula, C. R. (1971). Some suggestions for improving the psychology laboratory course experience. *American Psychologist, 26*(11), 1020–1021. https://doi.org/10.1037/h0032444

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*(1–3), 107–142. https://doi.org/10.1016/0010-0277(92)90041-f

Roelofs, A. (2018). A unified computational account of cumulative semantic, semantic blocking, and semantic distractor effects in picture naming. *Cognition, 172*, 59–72. https://doi.org/10.1016/j.cognition.2017.12.007

Sá-Leite, A. R., Fraga, I., & Comesaña, M. (2019). Grammatical gender processing in bilinguals: An analytic review. *Psychonomic Bulletin & Review, 26*(4), 1148–1173. https://doi.org/10.3758/s13423-019-01596-8

*Sá-Leite, A. R., Haro, J., Comesaña, M., & Fraga, I. (2021). Of beavers and tables: The role of animacy in the processing of grammatical gender within a picture-word interference task. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.661175

Sá-Leite, A. R., Luna, K., Fraga, I., & Comesaña, M. (2020). The gender congruency effect across languages in bilinguals: A meta-analysis. *Psychonomic Bulletin & Review, 27*(4), 677–693. https://doi.org/10.3758/s13423-019-01702-w

*Sá-Leite, A. R., Oliveira, H. M., Soares, A. P., Carreiras, M., & Comesaña, M. (2017). *Unpublished raw data on the gender (in)congruency effect with transparent Portuguese nouns*. University of Minho.

Sá-Leite, A. R., Tomaz, À., Hernández-Cabrera, J. A., Fraga, I., & Comesaña, M. (2020). *What a transparent romance language with a Germanic gender-determiner mapping tells us about gender retrieval: Insights from European Portuguese [manuscript submitted for publication]*. School of Psychology, University of Santiago de Compostela.

*Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language, 48*(1), 169–194. https://doi.org/10.1016/S0749-596X(02)00508-9

*Schiller, N. O., & Caramazza, A. (2006). Grammatical gender selection and the representation of morphemes: The production of Dutch diminutives. *Language & Cognitive Processes, 21*(7–8), 945–973. https://doi.org/10.1080/01690960600824344

*Schiller, N. O., & Costa, A. (2006). Different selection principles of freestanding and bound morphemes in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1201–1207. https://doi.org/10.1037/0278-7393.32.5.1201

*Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 841–850. https://doi.org/10.1037/0278-7393.19.4.841

Schriefers, H., Jescheniak, J. D., & Hantsch, A. (2002). Determiner selection in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 941–950. https://doi.org/10.1037/0278-7393.28.5.941

Schriefers, H., Jescheniak, J. D., & Hantsch, A. (2005). Selection of gender-marked morphemes in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 159–168. https://doi.org/10.1037/0278-7393.31.1.159

*Schriefers, H., & Teruel, E. (2000). Grammatical gender in noun phrase production: The gender interference effect in German. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1368–1377. https://doi.org/10.1037/0278-7393.26.6.1368

Spalek, K., & Schriefers, H. J. (2005). Dominance affects determiner selection in language production. *Journal of Memory and Language, 52*(1), 103–119. https://doi.org/10.1016/j.jml.2004.09.001

Starreveld, P., & La Heij, W. (2004). Phonological facilitation of grammatical gender retrieval. *Language & Cognitive Processes, 19*(6), 677–711. https://doi.org/10.1080/01690960444000061

*van Berkum, J. J. A. (1997). Syntactic processes in speech production: The retrieval of grammatical gender. *Cognition, 64*(2), 115–152. https://doi.org/10.1016/S0010-0277(97)00026-7

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*. https://doi.org/10.31234/osf.io/zcf8s

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin, 128*(3), 442–472. https://doi.org/10.1037/0033-2909.128.3.442

Vigliocco, G., Lauer, M., Damian, M., & Levelt, W. J. (2002). Semantic and syntactic forces in noun phrase production. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28*(1), 46–58. https://doi.org/10.1037//0278-7393.28.1.46

Wang, M., & Schiller, N. O. (2019). A review on grammatical gender agreement in speech production. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.02754