Language and Cognition (2025), 17, e34, 1–26
doi:10.1017/langcog.2025.3

ARTICLE

Feminine fox, not so feminine box: constraints on linguistic relativity effects for grammatical and conceptual gender

James Brand[1], Mikuláš Preininger[2], Adam Kříž[2] and Markéta Cehánková[2]

[1]Department of Psychology, Edge Hill University, Ormskirk, UK and [2]Faculty of Arts, Charles University, Prague, Czech Republic
Corresponding author: James Brand; Email: James.brand.ac@gmail.com

(Received 04 April 2024; Revised 27 December 2024; Accepted 03 January 2025)

Abstract

The influence of grammatical gender on conceptual representations of gender has proven to be a controversial topic in the linguistic relativity literature, with empirical evidence in support of the Linguistic Relativity Hypothesis being highly task and context-dependent, as well as being modulated by the type of items being investigated (animates/inanimates). In this paper, we take a megastudy approach in order to investigate differences in results based on explicit and implicit paradigms that modulate the role of language and gender in their design. We present analyses of three experiments focussing on participants (total N = 4,621) with a grammatically gendered L1 (Czech), a non-grammatically gendered L1 (English) and L1-Czech in L2-English, and on three distinct semantic categories – people, animals and inanimates (total N_{items} = 1,208). Our results indicate that the most reliable effects of grammatical gender influencing conceptual gender (outside of the domain of people) are observed for items representing animals, with Czech participants showing congruency effects in both explicit and implicit paradigms, even in their L2. The evidence for effects on inanimates is substantially weaker and is highly restrained to explicit tasks. We discuss these results in relation to the Linguistic Relativity Hypothesis and highlight important methodological considerations for future research.

Keywords: conceptual representation; gender processing; grammatical gender; linguistic relativity; whorf

1. Introduction

The extent to which language may influence thought is a key question for linguistics, psychology, cognitive science, and other interdisciplinary fields. The Linguistic Relativity Hypothesis (LRH) posits that the structure and vocabulary of a language play a role in shaping the way that we perceive and conceptualise the world around us

(Whorf, 1956). Empirical investigations into this topic have focused on how language influences thought by exploring a wide range of domains where the LRH can be tested, predominantly focussing on how linguistic labels may affect the processing of perceptual information (e.g. Lupyan, 2012). However, the importance of investigating linguistic features that go beyond the label, such as aspects of grammar, has been a particular focus for researchers looking to more comprehensibly assess the LRH (Lucy, 2016; Thierry, 2016).

Grammatical gender is one such morpho-syntactic property of language that has gained considerable attention in the literature. Grammatically gendered languages assign nouns a specific gender category, which is typically either masculine, feminine or neuter, but can expand to far many more (Corbett, 1991) and can be assigned to both animate and inanimate nouns based on semantic, morphological, or phonological classifications (Kramer, 2020). For researchers interested in the LRH, the primary focus is whether grammatical gender 'rubs off' onto conceptual gender, that is if a word is grammatically feminine/masculine, is there any increased attention to the perceived femininity/masculinity of the word's referent?

Empirical work on this specific area of the LRH has been wide-ranging in terms of the different methods used, languages investigated, and participants sampled. A recent systematic review by Samuel et al. (2019) looked at which parameters were more or less likely to provide support for the LRH. When there was empirical support for the LRH it normally came from experiments where gender was highly salient in the design, e.g. voice choice (where a female or male voice is assigned to an item, Sera et al., 2002) and sex assignment (where items are assigned a female or male sex, Belacchi & Cubelli, 2012). The explicit role of gender in such tasks has been highlighted as a possible confound, with participants potentially using grammatical gender in a conscious way, which could mean that responses are being guided strategically by accessing grammatical gender information purposefully (Almutrafi, 2015). This has led to a greater focus on implicit designs, where the strategic use of grammatical gender, or indeed language more generally, is restricted to unconscious processing (e.g. Boutonnet et al., 2012; Sato & Athanasopoulos, 2018; Sato et al., 2020). Thus, if experiments are going to provide more reliable evidence to test the LRH they would need to address the potential confounds introduced when gender and language are inherent to the task.

Another area of contention in the empirical results is whether there is strong support for the LRH in both animate and inanimate items. Samuel et al. (2019) again highlighted this as a parameter that gives more nuance to the nature of the LRH, with stronger support for animates in comparison to inanimates. For example, Vigliocco et al. (2005) demonstrated that Italian speakers exhibit an effect of grammatical gender in tasks involving similarity judgements and semantic substitution errors, but only for items referring to animals, whereas for inanimate items there was no effect (see also Imai et al., 2014; Ramos & Roberson, 2011; Saalbach et al., 2012). These results were framed in relation to the sex and gender hypothesis, which posits that the systematic relationship between grammatical gender and biological sex for items referring to humans (e.g. *učitel* is the Czech word for a male teacher and is grammatically masculine) can be extended to other sexuated items where the biological sex is less transparent, such as animals (e.g. *žába* is the Czech word for frog and is grammatically feminine, so would more likely enhance the salience of feminine properties). Nonetheless, Vigliocco et al.'s results provide no evidence that these effects extend to items which do not refer to sexuated entities

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

(e.g. most inanimate nouns) or to languages with more than two grammatical gender classes (e.g. when there is a neuter class).

Additional work that examines the representation of conceptual gender across the lexicon comes from large-scale norming studies. Such experiments aim to provide normative estimates of how feminine, neutral, or masculine a word is, simply by asking participants to rate those words using a Likert scale (e.g. Scott et al., 2019). Whilst not primarily focussed on questions related to the LRH (as was the case in the rating study by Konishi, 1993), there has been interest in whether such megastudy datasets can reveal any low-level relationships between grammatical and conceptual gender. For example, Vankrunkelsven et al. (2024) collected conceptual gender ratings for 24,000 Dutch words and descriptively analysed the ratings based on grammatical gender, finding very weak correlations. However, such an analysis did not take into account the different types of semantic categories of the words, e.g. people/animals/inanimates, or compare it to L1/L2 speakers of a non-grammatically gendered language, therefore only limited insights into the LRH have been reported from such datasets.

2. The present study

This paper looks at the LRH in relation to grammatical and conceptual gender by taking a megastudy approach with a focus on large samples of items and participants. Such an approach is relatively rare in the literature on grammatical and conceptual gender, but given that there is not a strong consensus in the empirical evidence, such an approach will provide substantial data on which inferences can be made. Across three experiments, we aim to investigate the extent to which explicit and implicit paradigms may lead to different outcomes related to the LRH, whilst also focussing on the types of stimuli where effects may or may not be observed. We sampled participants who have an L1 that is either grammatically gendered (specifically Czech) or non-grammatically gendered (specifically English), providing us with a comparison across the two languages. Crucially, we also collected data from L1 Czech participants who completed the experiments entirely in English, providing us with data that can be used to understand how the effects transfer to an L2 context.

Czech is a Slavic language that has three different grammatical gender classes – feminine, masculine, and neuter – which are marked morphologically for nouns, adjectives, verbs, some pronouns, and numerals, and in special cases adverbs. In this paper, we focus exclusively on nouns as the only open-class part-of-speech category whose grammatical gender is inherent to their form and not determined by syntactic agreement with another word. While some nouns whose nominative singular form ends with -a are often feminine and nouns ending with -o tend to be neuter (e.g. *žába* [frog] and *město* [city]), most of the time there is no clear systematic relation between the phonological form of a noun and its grammatical gender (e.g. *stůl* [table] is masculine and *sůl* [salt] is feminine).

In Experiment 1 we analyse data from a large-scale norming experiment, where written word stimuli are rated along a conceptual gender scale, in Czech and English. Experiment 2 extends this methodology by using images as the stimuli being rated, allowing for a direct comparison between the ratings for words and images. Experiment 3 uses a more implicit design inspired by Sato and Athanasopoulos (2018), where participants are briefly shown an image and are then asked to choose a face that

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

they associate the image more with, choosing between feminine, masculine, and neutral faces.

Across the experiments, our analyses investigate effects for three distinct categories of items – people, animals and inanimates. When considering the LRH, we would expect to find the following patterns in the data:

(1) People – No differences across languages. Animate nouns referring to people will have conceptual gender that is driven more by biological sex or cultural stereotypes.
(2) Animals – Czech participants rating Czech words and Czech participants rating English words will have an effect on grammatical gender, whereby grammatically feminine/masculine words will be conceptually more feminine/masculine. Whereas English participants rating English words will not have any congruency with the grammatical gender of the given word in Czech in terms of their conceptual ratings.
(3) Inanimates – The same pattern as described for animals, but the effect will be substantially smaller, due to the word referring to an inanimate noun, which would generally not carry any information about biological sex.

3. Experiment 1: Rating written words

The primary goal of this experiment was to explore whether ratings of conceptual gender differ between participants from three different conditions: L1 Czech participants rating Czech words, L1 Czech – L2 English participants rating translation equivalents of these words in English, and L1 monolingual English participants also rating the translation equivalents in English. We aimed to examine the following research question: Does the grammatical gender of the word in Czech lead to differences in the ratings across the participant groups, specifically when the words refer to (1) people, (2) animals and (3) inanimates?

3.1. Methods

3.1.1. Participants

Summary information on the final sample (that is after filtering of participants) across the different conditions is provided in Table 1.

Czech sample. Data reported for this sample comes from a larger project reported in Preininger et al. (2022; 2024), we use a subset of this data, which is described below.

The participants who completed the rating study in Czech were predominately students from Charles University in the Czech Republic. They were from a university-wide participant pool and completed the experiment for course credit. We removed 48 participants from this sample who reported that they were not a native speaker of Czech or if they reported any additional languages as their native language, e.g. those who considered themselves raised as bilingual. We also removed data from 4 participants who completed the same version of the experiment more than once, keeping only the data from their initial completion. This left us with 1,223 participants.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

Table 1. Summary of demographic information for the final sample of participants in Experiments 1, 2 and 3

	n Total	n Female	n Male	n Non-binary	n Other	Median age
Experiment 1						
Czech	1423	1050	367	2	4	22
CzEng	990	793	183	11	3	21
English	100	68	25	7	0	21
Experiment 2						
Czech						
Words	233	192	40	0	1	22
Colour	240	201	35	4	0	21
Gray	244	199	42	3	0	21
CzEng						
Words	189	153	35	0	1	21
Colour	146	119	26	1	0	21.5
Gray	149	124	24	1	0	21
English						
Words	102	70	25	7	0	21
Colour	99	73	24	2	0	22
Gray	100	74	26	0	0	21
Experiment 3						
CzEng	366	282	78	5	1	21
English	240	144	92	4	0	22

We also recruited participants from Prolific (www.prolific.com) (n = 200), this was to collect more participants who identified as male, as there was a large bias for female participants from our student sample. All participants from this subsample declared that their native language was Czech and that they were located in the Czech Republic. They were paid £4.28 for completing the experiment.

L1Czech-L2English sample (CzEng). The participants who completed the rating study in English, but had an L1 of Czech, were students from Charles University in the Czech Republic. They were from a university-wide participant pool and completed the experiment for course credit. We removed 79 participants from this sample who reported that they were not a native speaker of Czech or reported any additional languages as their native language, e.g., those who considered themselves raised as bilingual. The final number of participants for this sample was 990.

L1 English sample. The participants who completed the rating study in English, and who had an L1 of English were recruited from Prolific. They reported that they were monolingual and born in the UK, so it is assumed that they were British English speakers. We also filtered participants so that the current student status was set to true. They were paid £3.38 for completing the experiment. The final number of participants for this sample was 100.

3.1.2. Stimuli

Our stimuli for this experiment comprised exclusively of orthographically presented word forms. Below, we will outline the word lists for Czech and English items, with the final word list used in analysis and the filtering criteria also described. The final

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

word list for Experiment 1 contained 800 words, which were grammatically masculine or feminine in Czech and for which we had ratings from all the participant groups (Czech, CzEng and English).

Czech words. We initially presented a total of 2,999 items to participants in the Czech condition. The word list contained nouns, adjectives and verbs and spanned a wide range of semantic categories (e.g. role nouns, tools, food, emotions, politics) and corpus-based frequencies (based on the Syn-v9 corpus of written Czech, see Křen et al., 2021). There were 482 words taken from the Czech responses provided in the Multilingual Picture Database (Duñabeitia et al., 2022), where standardised names for images were collected. The other words were from the word list reported by Preininger et al. (2022). When possible, for role nouns and adjectives, both the grammatically feminine and masculine forms were included, e.g. the adjective '*polite*' was included as both '*zdvořilá*' (fem) and '*zdvořilý*' (masc), or the role noun '*mechanic*' was included as both '*mechanička*' (fem) and '*mechanik*' (masc).

English *words*. We presented to the CzEng participants translation equivalents of all the Czech words. Translations were produced and agreed upon by three native Czech speakers, all of whom were also highly proficient speakers of English. This resulted in a list of 2,874 unique words. This number is smaller than the Czech list as adjectives could only be given as a single form, e.g. '*polite*' was the only translation of the two gender variants presented in Czech. For role nouns, we presented the form in a way that explicitly marks the gender of the referent, e.g. mechanic was presented as both '*mechanic (female)*' and '*mechanic (male)*', as well as the unmarked version, e.g. '*mechanic*'. For words that could be interpreted as ambiguous, we added a word in parentheses to ensure the meaning was aligned to the Czech word, e.g. '*rubber*' was presented as '*rubber (eraser)*'. Within this list were 548 unique words taken from the British English naming data from the Multilingual Picture Database. Due to the cost of recruitment of native speakers of English, we were only able to collect data for this subset of words from the English participants, and not the complete list of words presented to the CzEng participants.

To increase the number of items for which we had data from native English speakers for English words in our list, we used two additional datasets that had ratings for conceptual gender – the Glasgow Norms (Scott et al., 2019) and data reported in Lewis et al. (2022). There were 745 words from these two datasets that were in our English word list, but we did not have data for them (Glasgow: n = 690; Lewis et al: n = 55). When a word occurred in both datasets, preference was given for the Glasgow Norms values, as they were from UK-based participants and the design was most similar to ours. We transformed the mean ratings to align with the format of our existing data (see supplementary materials and analysis below). To check that the data was comparable to our existing English data, we ran a Pearson's correlation on mean ratings for words that appeared in our data and the additional datasets (Glasgow: $r(387) = .935$, $p < .001$; Lewis et al: $r(214) = .873$, $p < .001$). We were therefore confident that these extra ratings were comparable to our original data.

Final word list. We compiled all the words from the Czech and English lists together. We coded grammatical gender manually and cross-referenced an online dictionary when unsure (https://prirucka.ujc.cas.cz/). We then further manually coded each

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

Table 2. Summary of counts for items based on the grammatical gender in Czech for stimuli analysed in Experiments 1, 2, and 3

	Grammatically feminine	Grammatically masculine	Total
Experiment 1			
Animal	34	51	85
Inanimate noun	358	286	644
Person	24	47	71
Experiment 2 & 3			
Animal	28	43	71
Inanimate noun	148	150	298
Person	7	32	39

item for animacy (animate/inanimate), part of speech (adjective/noun/verb) and item category (animal/ inanimate object/person/other). We then filtered the list so that only words that were either grammatically feminine or masculine in Czech (and their translation equivalents in English) remained. This left 800 items where we had both Czech and English words. See Table 2 for the counts of items in each grammatical gender and semantic category.

3.1.3. Procedure

Participants took part in a rating experiment as described in Preininger et al. (2022), where participants rated written word forms along 7-point Likert scales for five different socio-semantic dimensions, that is gender, age, location, political alignment, and valence. The experiments were presented entirely in Czech for the Czech sample and entirely in English for the CzEng and English samples.

For example, for the gender dimension, participants were instructed to rate the words based on "*The degree to which you associate the word meaning with masculinity or femininity.*" The Likert scale was anchored with '*very masculine*' on the left, '*neutral*' in the centre and '*very feminine*' on the right, with the choice of '*I do not know the word*' as an additional option. Words were divided into lists of 105 to 114 items, which were presented in a randomised order, with a calibrator word of '*necklace/náhrdelník*' and 5 pseudowords which acted as controls. When there was the possibility of having a grammatically feminine and masculine form, (e.g. for adjectives in Czech), we ensured that each list only contained one variant, with the other being presented in a separate list. The majority of participants only completed one list of items, however, there was a small number who completed multiple lists. Our instructions made it clear that participants did not have to respond under any time pressures. See Figure 1 for an example of the gender dimension and the supplementary materials for an example of the full experiments in each language.

3.2. Statistical analysis

We conducted all analyses using R version 4.3.1. (R Core Team, 2023). All data, preprocessing and analysis scripts, as well as package versions, are available at https://osf.io/uky6q

Across the 800 items, the mean proportion of participants who knew a word was very high (Czech: 0.999, CzEng: 0.963, English: 0.999) and the median number of

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

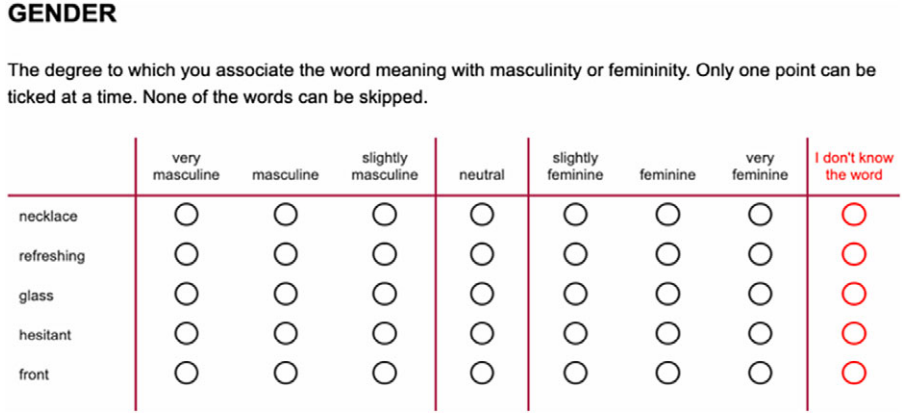

Figure 1. Example of the word rating experiment in English for Experiment 1, 'necklace' is the calibrator word that was always presented as the first word.

responses per word was also high (Czech: 46, CzEng: 36, English: 20). We converted the ratings to a numeric scale[1] (−3, −2, −1, 0, 1, 2, 3) and then calculated the mean from these values for each item. When comparing the ratings from all the items, across the three languages, there was a normal distribution centred around the neutral midpoint of 0 (Czech: $M = -.064$, $SD = .914$, CzEng: $M = -.084$, $SD = .930$, English: $M = -.054$, $SD = 1.018$).

We analysed the resulting dataset using linear regression models through the lm function. We first ran a model predicting the mean conceptual gender rating as the dependent variable by a 3-way interaction term of Czech_grammatical_gender (feminine/masculine)*language_group(Czech/CzEng/English)*item_category(person/animal/inanimate noun). Any significant interactions from the model were first assessed using the ANOVA function for model comparison. If there was a significant interaction, we ran pairwise comparisons to assess differences between grammatical genders in each of the three language groups and to compare differences between language groups for grammatically feminine and masculine items, correcting for multiple comparisons with Tukey-adjusted significance testing, this was done using the emmeans package.

3.3. Results and discussion

The model had a significant 3-way interaction term ($F(4, 2382) = 3.028$, $p = .017$, $R^2 = .389$). The data are visualised in Figure 2. Below, we report the results from pairwise comparisons for items referring to people, animals, and inanimate nouns.

[1]As Liddell and Kruschke (2018) and Taylor et al. (2022) highlight, using a mean for ordinal rating data is not an optimal approach for calculating norms. However, as we are using secondary data to supplement our English ratings – where only the mean is available and not all the raw rating data – we unfortunately could not run the more suitable analyses. As a result, our data will conform to the more traditional approach of using item-level means across all our data. We address this point in Experiment 2 by using only raw data and analyses suited to ordinal data.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

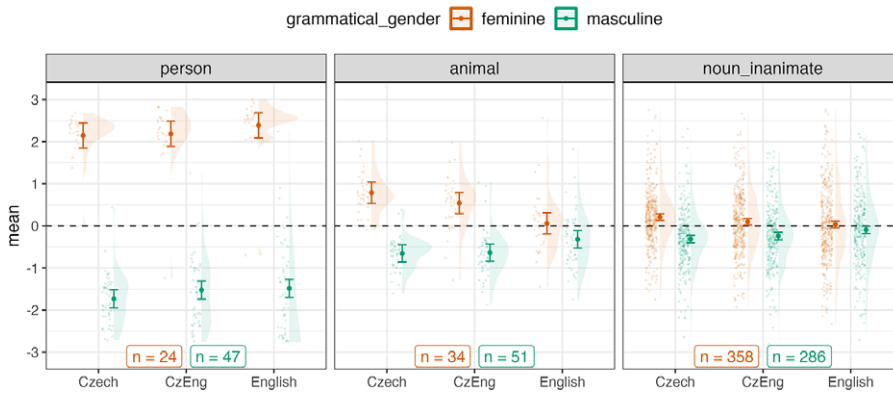

Figure 2. Visualisation of the data from Experiment 1. Facets are used for each of the separate categories analysed. The x-axis is used for the different language groups and the y-axis represents the mean conceptual gender, with positive values for feminine associations and negative values for masculine associations. The dashed line represents the neutral midpoint. Model estimates and 95% confidence intervals are given as solid colours, with the lighter points representing individual words. Orange is used for grammatically feminine items, green for grammatically masculine based on Czech grammatical gender. The number of items in each grammatical gender are given in the 'n = x' labels.

3.3.1. People

For the pairwise comparisons, there were significant effects for all three language groups based on the comparison between grammatical genders (Czech: $b = 3.875$, $t = 20.612$, $p < .0001$; CzEng: $b = 3.711$, $t = 19.738$, $p < .0001$; English: $b = 3.873$, $t = 20.598$, $p < .0001$). There were no significant differences when comparing language groups within grammatical genders for either feminine or masculine items (all p's $> .05$).

This indicates that all three language groups exhibit an effect where conceptual and grammatical gender are related in a congruent way. It also indicates that there were no differences in ratings across the languages for grammatically feminine or masculine items. Although not surprising, we interpret these results as evidence that grammatical gender and conceptual gender are aligned systematically for these items, either through biological sex (e.g. 'King/Král' refers to males and is grammatically masculine in Czech) or cultural stereotypes (e.g. 'caregiver/pečovatelka' is stereotypically associated to females and is grammatically feminine in Czech). The knowledge of grammatical gender for Czech and CzEng participants does not result in any clear differences in ratings when compared to the English participants.

3.3.2. Animals

For the pairwise comparisons, there were significant effects for all three languages based on the comparison between grammatical genders (Czech: $b = 1.439$, $t = 8.673$, $p < .001$; CzEng: $b = 1.175$, $t = 7.084$, $p < .001$; English: $b = .378$, $t = 2.280$, $p = .023$). There were no significant differences when comparing Czech and CzEng groups for grammatically feminine or masculine items (all p's $> .05$). When comparing Czech and CzEng to English, we observed significant differences for grammatically feminine items (Czech-English: $b = .727$, $t = 3.998$, $p < .001$; CzEng-English: $b = .481$, $t = 2.646$, $p = .022$), but only marginal effects for grammatically masculine items

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

(Czech-English: $b = -.344$, $t = -2.251$, $p = .063$; CzEng-English: $b = -.316$, $t = -2.131$, $p = .084$).

Although it is not conclusive, we could interpret these results as potential evidence in support of the LRH. Animals differ from people as they do not normally have a clear biological association to gender, e.g. it is not clear if a wasp is biologically female or male. The ratings from Czech and CzEng participants are not distinguishable from each other, even though the words were rated in different languages. English participants, however, also showed the same association towards the grammatical gender in Czech, but to a much smaller extent. It is possible that there is, for some items, a systematic relationship between the conceptual and grammatical gender in Czech, e.g. *dog/pes* is conceptually masculine for all language groups and is grammatically masculine in Czech. Nevertheless, it appears that the Czech and CzEng participants, who have an explicit knowledge of the Czech grammatical gender, show a much larger congruency effect. This could be explained in relation to the LRH, with the grammatical gender contributing to the conceptual representation of words for animals in both Czech and CzEng. The results also indicate that grammatically feminine animals appear to be contributing more to this effect compared to grammatically masculine animals, but it is not clear why this might be.

3.3.3. *Inanimate nouns*

For the pairwise comparisons, we observed all languages exhibiting a significant difference between grammatically feminine and masculine items (Czech: $b = .523$, $t = 8.803$, $p < .001$; CzEng: $b = .338$, $t = 5.693$, $p < .001$; English: $b = .130$, $t = 2.181$, $p = .029$). This indicates that all three language groups demonstrated a trend for ratings to be congruent with Czech grammatical gender. Although these effects were significant, the magnitude of the estimates was not similar – the largest difference was found for Czech, the smallest for English. Looking at the differences between languages, we observed no differences between Czech and CzEng (both p's > .05). Czech differed significantly from English for both feminine ($b = .177$, $t = 3.157$, $p = .005$) and masculine ($b = -.217$, $t = -3.458$, $p = .002$) items. For CzEng and English, there was no effect for feminine items ($b = .065$, $t = 1.161$, $p = .477$) and only a marginal effect for masculine items ($b = -.144$, $t = -2.293$, $p = .057$).

Interpreting these results as clear evidence that supports the LRH is not straightforward. As we observed for animals, we observed effects across all language groups, even for English, which demonstrated that participants are rating the words in a way that is congruent with the grammatical gender of Czech. However, there is not a clear indication that the CzEng and English participants are rating the words differently, whereas the Czech participants do appear to differ from the English participants. This could indicate that when using written stimuli and explicit rating judgements, the morphological marking of grammatical gender in Czech may be contributing to the way the ratings are made. It is also worth noting that the size of the estimates, although significant across all languages, is relatively small. Given the relatively large number of items in our experiment, there may have been systematic relationships between conceptual and grammatical gender that are not accounted for in our statistical modelling, e.g., '*moustache/knír*' is associated with masculinity across the languages, but is also grammatically masculine in Czech, which could explain the significant effect found in English.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

4. Experiment 2: Rating image stimuli

The primary aim of Experiment 2 was to extend the results reported in Experiment 1, but this time with stimuli that were images and not words. The motivation for this is that images do not have explicit linguistic information, and the activation of the word label is not as prominent (Levelt, 1989), therefore reducing any explicit linguistic effects that may have influenced the results from Experiment 1. We aimed to address 2 main research questions. (1) Do we find the same pattern of effects for items that are related to people, animals, and inanimate nouns as we observed in Experiment 1 in image stimuli? (2) Are there differences in these effects across words, colour images and grayscale images?

4.1. Methods

4.1.1. Participants

The summary of the participant demographics for the final sample analysed is presented in Table 1.

For the word ratings, we used a subset of the data from Experiment 1, where only the words coming from the Multilingual Picture Database were kept for analysis. For the image ratings, we collected data from a new set of participants. For the Czech and CzEng samples, we recruited participants from the university-wide participant pool at Charles University in the Czech Republic, with all participants receiving course credit. We collected data from 242 participants for the Czech data on colour images (excluding 2 non-native Czech speakers) and 253 participants for the grayscale images (excluding 9 non-native Czech speakers). We collected data from 156 participants for the CzEng data on colour images, (excluding 10 non-native Czech speakers) and 164 participants for the grayscale images (excluding 15 non-native Czech speakers). For the English sample, we again recruited participants from Prolific who were located in the UK and were monolingual English speakers. We also filtered participants so that the current student status was set to true. They were paid £3.38 for completing the experiment. We collected data from 99 participants for the colour images and 100 for the grayscale images.

4.1.2. Stimuli

We used three different sources of stimuli – words, colour images and grayscale images. The stimuli were taken from the Multilingual Picture Database, where naming data for images was standardised across participants from different languages for a set of 500 colour images. From this database, we were able to determine the word used to name each of the images in Czech and British English. The database has colour and grayscale versions, which will be the basis of our image stimuli, whilst the word labels will act as the word stimuli.

We again manually coded each of the words for grammatical gender in Czech and item categories (person, animal, or inanimate noun), assigning these values to each of the Czech and English words and image stimuli. We kept only items that were grammatically feminine or masculine in Czech. We excluded a further 21 items which had the same meaning, e.g. there are two pictures that have the name 'doctor', ensuring that all images had only one unique label in Czech and English. This left 408 items in each of the stimuli conditions. See Table 2 for a summary.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

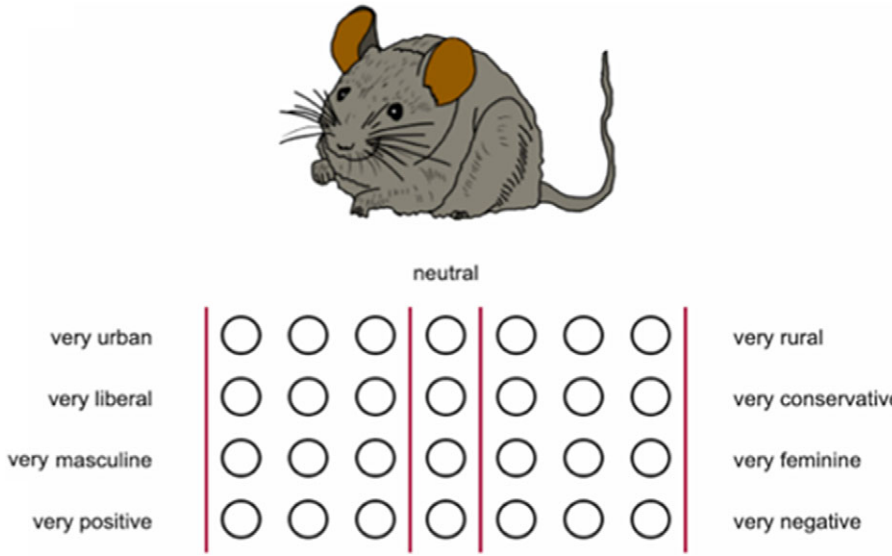

Figure 3. Example of the colour picture rating experiment in English for Experiment 2.

4.1.3. Procedure

The data collected for word stimuli came from Experiment 1. For the image stimuli, the procedure was similar to that used in Experiment 1, but participants saw the image displayed on the screen and the rating scales were presented directly below the image, see Figure 3 for an example. The calibrator image, which was presented as the first item to be rated was a picture of a mouse. Participants rated 101 image stimuli during the experiment in a randomised order. After rating all the images, participants also completed an additional rating task, which is described in the stimuli section of Experiment 3. Once again, participants were instructed that they do not have to respond under any time pressures.

4.2. Statistical analysis

To obtain estimates of the normative ratings for each item, we ran cumulative mixed-effects models on the raw data using the ordinal packages in R (Christensen, 2018), see Taylor et al. (2023). This approach was preferred to using the mean as they allow for the data to be modelled as an ordinal variable (which the Likert scales were) and they can take into account random sources of variation. Thus, we coded the ratings as ordinal values ($-3 < -2 < -1 < 0 < 1 < 2 < 3$), then predicted them using the clmm function, with a random intercept for the participant and a non-correlated random slope for version (word/colour/grayscale) on the item intercepts. We ran individual models for each language group, extracting the random intercepts for each slope. We then used these values in a regression model, with a 4-way interaction term between grammatical gender (feminine/masculine), language (Czech/CzEng/English), item category (person/animal/inanimate noun) and stimuli type (word/colour/grayscale).

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

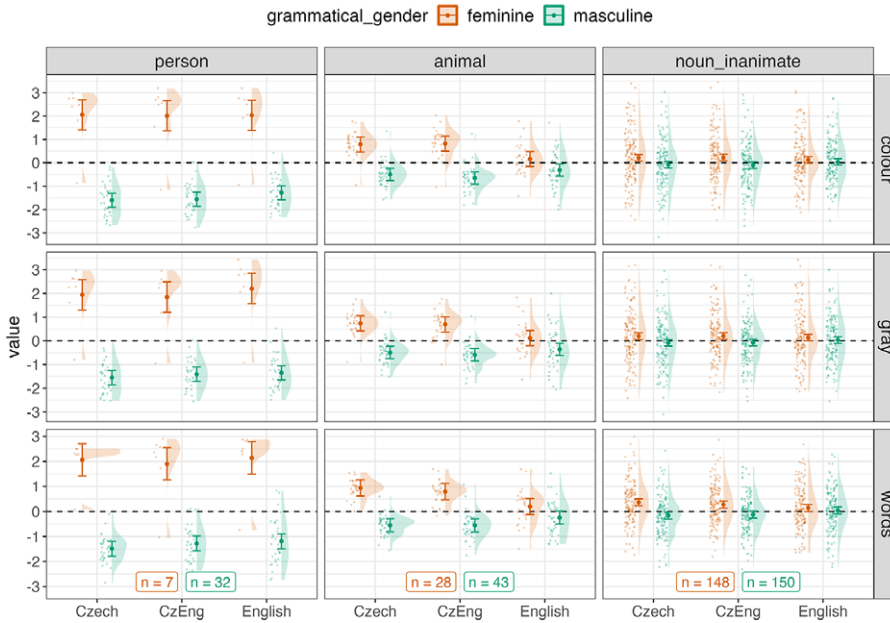

Figure 4. Visualisation of the data from Experiment 2. Facets at the top are used for each of the separate item categories analysed, facets on the right are for the stimuli types. The x-axis is used for the different language groups and the y-axis represents the mean conceptual gender, with positive values for feminine associations and negative values for masculine associations. The dashed line represents the neutral midpoint. Model estimates and 95% confidence intervals are given as solid colours, with the lighter points representing individual items. Orange is used for grammatically feminine items, green for grammatically masculine based on Czech grammatical gender. The number of items in each grammatical gender are given in the 'n = x' labels.

4.3. Results and discussion

The model did not have a significant 4-way interaction term ($F(8, 3618) = .091$, $p = .999$), but there was a significant 3-way interaction term between grammatical_gender*language*item_category ($F(4, 3618) = 4.160$, $p = .002$, $R^2 = .290$). The data are visualised in Figure 4. Below, we report the results from pairwise comparisons for people, animals, and inanimate nouns; summaries are provided in Table 3.

4.3.1. People

For the pairwise comparisons, there were significant effects for all three languages based on the comparison between grammatical genders, with significant effects in all stimuli types (all p's < .001). There were no significant differences when comparing language groups within grammatical genders for either feminine or masculine items; this pattern was observed regardless of stimuli type (all p's > .05).

This indicates that for items referring to or depicting people, there were no differences across all languages and stimuli types, where conceptual and grammatical gender are related in a congruent way. Thus, there is no evidence that participants

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

Table 3. Pairwise comparisons for ratings of people, animals, and inanimate nouns in Experiment 2. The comparisons test the effect of grammatical gender on ratings for the different stimuli types for Czech, CzEng and English

	Estimate	*z*	*P*
Experiment 3: People			
CzEng			
Feminine-masculine	2.808	5.719	< .001
Feminine-neutral	1.967	5.237	< .001
Masculine-neutral	1.834	4.950	< .001
English			
Feminine-masculine	3.026	5.777	< .001
Feminine-neutral	1.838	4.613	< .001
Masculine-neutral	1.907	4.807	< .001
Experiment 3: Animals			
CzEng			
Feminine-masculine	1.147	4.173	< .001
Feminine-neutral	.599	2.853	.004
Masculine-neutral	.817	3.904	< .001
English			
Feminine-masculine	.562	1.991	.046
Feminine-neutral	.259	1.172	.241
Masculine-neutral	.442	2.020	.043
Experiment 3: Inanimate nouns			
CzEng			
Feminine-masculine	.149	1.135	.256
Feminine-neutral	.144	1.427	.153
Masculine-neutral	.158	1.581	.114
English			
Feminine-masculine	.177	1.299	.194
Feminine-neutral	.037	.348	.728
Masculine-neutral	.187	1.774	.076

who have a grammatically gendered language as their L1 (Czech and CzEng) differ from those participants who do not, as was the case in Experiment 1.

4.3.2. Animals

For the pairwise comparisons, there were significant effects for all three languages based on the comparison between grammatical genders, with significant effects in all stimuli types (all p's < .05). There were significant differences when comparing language groups for grammatically feminine items, with both Czech and CzEng differing from English for all stimuli types (all p's < .05), but there were no significant differences for grammatically masculine items (all p's > .05). Czech and CzEng had no significant differences for either grammatically feminine or masculine items.[2]

As was the case in Experiment 1, we again observed potential support for the LRH, in both words and images, with Czech and CzEng participants rating items in a way that is congruent with their grammatical gender in Czech. Specifically, animals that are grammatically feminine are conceptually more feminine in both Czech and CzEng conditions, when compared to English participants. However, we did not observe similar patterns for grammatically masculine items. It should also be noted

[2]Note that these results do not differ when animals that might have a specific word for female/male sexes are included (n = 6), e.g. an image of a lion with a mane.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

that the size of the estimates for both Czech and CzEng were substantially larger (~1.3) in comparison to those for English (~0.5), indicating that the effect is much more prominent for those participants with Czech as their L1, which may indicate a role for grammatical gender contributing to the effect.

4.3.3. Inanimate nouns

For the pairwise comparisons, there were significant effects for Czech and CzEng based on the comparison between grammatical genders, with significant effects in all stimuli types (all p's < .01), as was the case in Experiment 1. However, there were no significant effects of grammatical gender for the English participants in any of the stimuli types (all p's > .05). However, the comparisons across language groups for grammatically feminine and masculine items were all not significant (all p's > .05).

The presence of a grammatical gender effect in Czech and CzEng, along with a lack of an effect in English, could be interpreted as tentative evidence in support of the LRH. However, we would highlight the fact that there were no across-language differences, that is for grammatically feminine and masculine ratings, there was no evidence to suggest the actual ratings differed between Czech, CzEng and English participants. Furthermore, we also note that the estimates for the grammatical gender effects reported in Czech and CzEng were relatively small (~0.3 for image stimuli, ~0.45 for words), suggesting that any contribution of grammatical gender from the participant's L1 is very subtle for inanimate nouns.

5. Experiment 3: Image-face decision experiment

The primary aim of Experiment 3 is to expand on the results from Experiments 1 and 2. Specifically, by investigating whether CzEng and English participants display a similar pattern of results, but within a more implicit experimental paradigm. To achieve this, we adapted the experimental design used by Sato and Athanasopoulos (2018), whereby we presented individual image primes and asked participants to make a timed decision between two faces in either feminine-masculine, feminine-neutral or masculine-neutral conditions; thus, we do not use the explicit rating scales from the previous experiments. Whilst the feminine-masculine condition still holds with our overarching hypotheses, the feminine-neutral and masculine-neutral conditions act as a test for the same hypotheses, but in a way where the gender distinction is less salient.

5.1. Methods

5.1.1. Participants

The summary of the participant demographics for the final sample analysed is presented in Table 1. Note that we only collected data from CzEng and English language groups as the experiment was purposefully non-linguistic; that is, the main trials contained no written stimuli. Thus, we did not believe that collecting another large sample of data (where the only difference was the language used in the instructions) would be beneficial.

CzEng. We again recruited participants for the CzEng group from the university-wide participant pool at Charles University in the Czech Republic. All participants

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

were given course credit for taking part. Our initial sample was from 424 participants, but we removed 36 non-native Czech participants and 15 participants who reported that English was not their dominant L2. A further 7 participants were removed based on an inspection of their median reaction times during the experimental trials, whereby participants who had a median reaction time smaller than 300 ms or larger than 3000 ms were excluded on the basis that they were either considerably faster or slower in comparison to all other participants. This left us with 366 participants.

English. We again recruited participants for the English group from Prolific who were located in the UK and were monolingual English speakers. We also filtered participants so that the current student status was set to true. Participants were paid £1.75 for completing the experiment. Our initial sample was from 241 participants, but we removed 1 participant on the basis of their median reaction time being larger than 3000 ms during experimental trials.

5.1.2. Stimuli

The images that were used as items in this experiment were the same 500 images from the Multilingual Picture Database that were used in Experiment 2, but we only presented the grayscale images. This decision was based on the fact that there were no major differences in the colour and grayscale ratings from Experiment 2 and that the grayscale versions will remove any underlying bias between conceptual gender and colour, e.g. a pink dressing gown will be conceptually more gender laden than a grayscale version. The 500 images were divided into 4 different lists, each containing 125 items. Across the lists, there were an approximately equal number of grammatically feminine (44–47), masculine (55–58), and neuter/ambiguous (21–24) items.

The stimuli used for the faces came from images generated using FaceGen Modeler (Singular Inversions Inc.). We initially generated 10 female, 10 male and 12 neutral faces, all had neutral expressions and were masked so only the face was visible (no hair, ears or neck were visible). The resulting stimuli were rated by all participants at the end of the image rating procedure in Experiment 2, following the exact same procedure, that is rating the faces along different dimensions, including gender for colour and grayscale versions. We then subsequently chose 6 faces from each of the female, male and neutral conditions based on the mean grayscale ratings from CzEng (n = 149) and English (n = 100) participants, resulting in 18 distinct face stimuli. For the female faces, there was very little variation across items so the final 6 faces were selected randomly and were all rated as conceptually feminine (M = 2.213, SD = .087), this was also the case for the male faces, which were all rated as conceptually masculine (M = −2.197, SD = .104). For the neutral faces, there was substantially more variation across items, so we chose the 6 faces with ratings closest to the midpoint of the scale, which represented conceptually neutral gender (M = −.097, SD = .300). See Figure 5 for a visualisation of the summary data from all the face ratings.

5.1.3. Procedure

The experiment was programmed in JsPsych 7.3 (de Leeuw et al., 2023) and hosted on a JATOS server (Lange et al., 2015). The experiment was presented entirely in English for both the CzEng and English participants.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

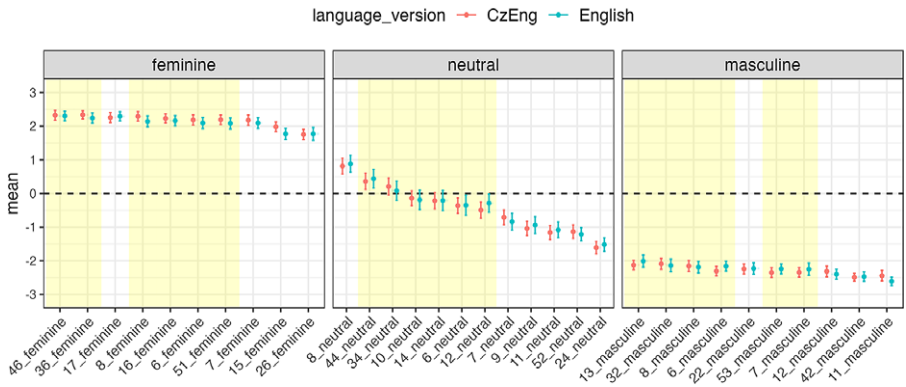

Figure 5. Visualisation of the norming data for face stimuli in Experiment 3. Facets are used for each of the gender categories. The x-axis is used for the different faces and the y-axis represents the mean conceptual gender, with positive values for feminine ratings and negative values for masculine ratings. The dashed line represents the neutral midpoint. Means and 95% confidence intervals are given for CzEng and English ratings. The items highlighted in yellow were used for Experiment 3.

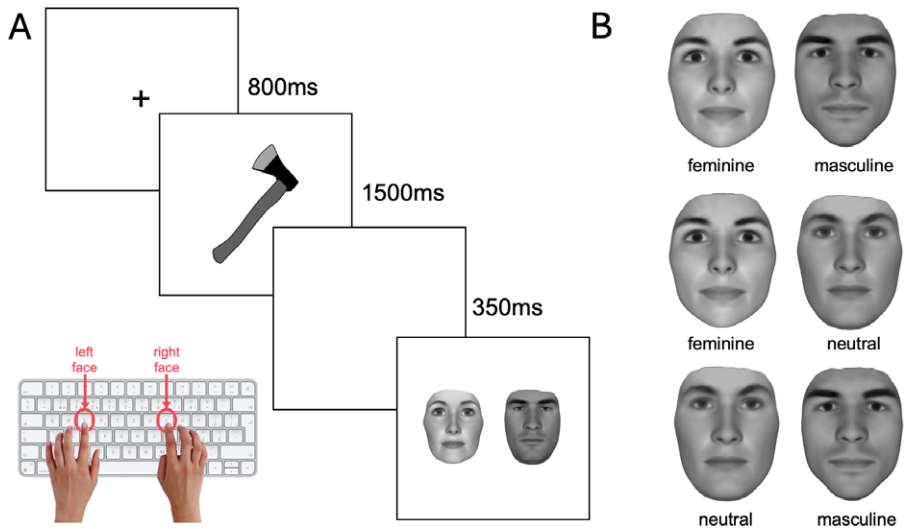

Figure 6. A: Visualisation of the procedure used for each trial in Experiment 3. B: Examples of the three types of face combinations presented to participants.

Participants were presented with a fixation cross in the middle of the screen for 800 ms, followed by randomly selected image stimuli for 1500 ms, and then a blank screen for 350 ms. Immediately after the blank screen, two faces appeared, one on the left and one on the right side of the screen, see Figure 6A. The trial ended when the participant pressed either the 'e' or the 'i' key. There were 2 practice trials at the beginning so that the participants could familiarise themselves with the procedure, the items shown were an axe and a heel of a shoe, then each participant completed 125 trials, with the chance to take a break halfway through. Participants were

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

provided with the following instructions: "*If you associate the picture more with the face on the left, please press the "e" key on your keyboard, and if you associate the picture more with the face on the right, please press the "i" key. Please respond as quickly as possible, but make sure you are paying attention to the picture and both of the faces whilst completing the experiment. In case you are not sure about which face to choose, simply go with your first instinct, there are no correct answers* "

When responding to the faces, there were 3 possible presentation combinations: i) feminine-masculine, ii) feminine-neutral, or iii) neutral-masculine, see Figure 6B for examples of each. The number of face combination trials was distributed approximately equally across the experiment, with either 41 or 42 trials per combination. The location of the faces (that is left or right on the screen) was randomised per trial for each participant. Each of the 18 faces was presented either 13 or 14 times throughout the experiment. Participants would not encounter the same image prime more than once during the experiment. Each image prime was responded to in each of the 3 face combination conditions by different participants (median n: CzEng = 30.5, English = 20). This means that for any given image prime there would be an average of 20 English participants who responded to the feminine-masculine trial, a separate 20 participants for the feminine-neutral trial and another separate 20 participants for the neutral-masculine trial.

5.2. Statistical analyses

Before analysing the data, we first removed any items where the word label was not grammatically feminine or masculine, leaving the same 408 items used in Experiment 2. The mean reaction time for participants to choose a face for these items was 1084 ms (IQR = 718-1651 ms), calculated from 61,803 trials. We then trimmed the data so that the quickest 2.5% (<343 ms, 1,535 responses) and the slowest 2.5% (>3892 ms, 1,545 responses) of the data were removed. This was on the basis that these items would have been extremely quick or slow, relative to all other responses. The resulting data comprised 58,723 trials.

We used the same coding for the items as used in Experiment 2, where the item category was categorised as depicting people, animals, or inanimate nouns, language group as CzEng or English and grammatical gender as feminine or masculine. We then coded participant responses to each of the face combination trials as binary values. For feminine-masculine trials, a response for the feminine face was coded as 1, and the masculine face as 0. For the feminine-neutral trials, feminine is 1 and neutral is 0. For the neutral-masculine trials, neutral is 1 and masculine is 0. We analysed the data using a generalized linear mixed-effects model using the lme4 package in R (Bates et al., 2015). Random intercepts for participant and item (with a random slope for face combination) were included, and the bobqa optimizer was used to resolve convergence warnings. We ran follow-up pairwise comparisons to test whether there were effects of grammatical gender within each language group for the different item categories, in addition to testing whether there were differences across language groups.

5.3. Results and discussion

Our model predicted the binary response variable initially with a 3-way interaction term of language_group*grammatical_gender*item_category, but a likelihood ratio

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

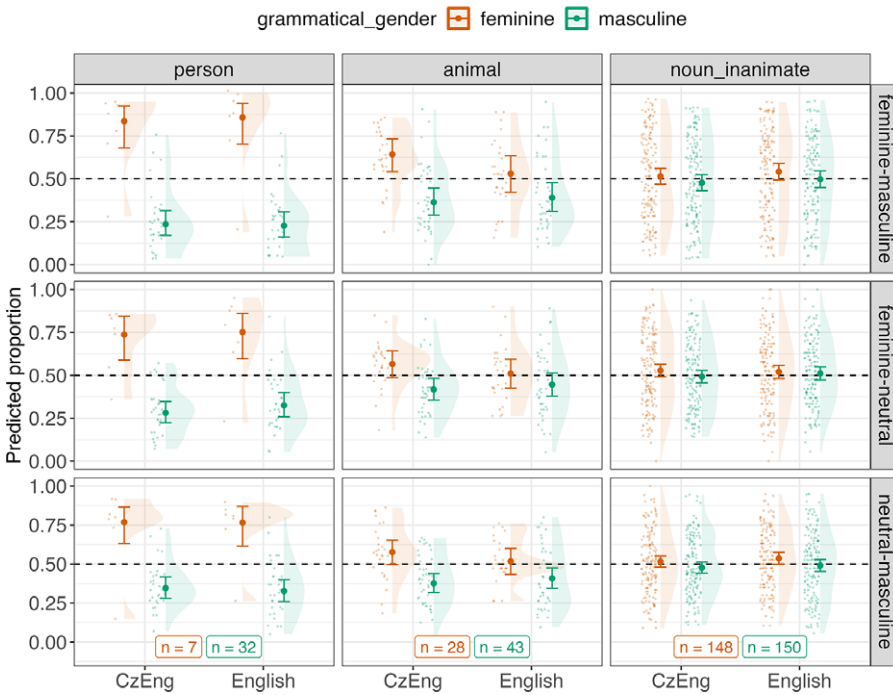

Figure 7. Visualisation of the data from Experiment 3. Facets at the top are used for each of the separate item categories analysed, facets on the right are for the different face combination conditions. The x-axis is used for the different language groups and the y-axis represents the predicted proportion of choosing the less masculine face, with values closer to 1 for feminine faces in the feminine-masculine and feminine-neutral conditions, or neutral faces for the neutral-masculine condition. The dashed line represents the 0.5 midpoint, where the proportion of responses to each face is equal. Model estimates and 95% confidence intervals are given as solid colours, with the lighter points representing individual items. Orange is used for grammatically feminine items, green for grammatically masculine based on Czech grammatical gender. The number of items in each grammatical gender are given in the 'n = x' labels.

test revealed improved model fit by including a 4-way interaction term of language_group*grammatical_gender*item_category*face_combination ($\chi^2(22) = 70.47$, $p < 0.001$). The data are visualised in Figure 7. Below, we report the results from pairwise comparisons for people, animals, and inanimate nouns, summaries are provided in Table 4.

5.3.1. People

For the analyses containing only items that depict people, we observed significant effects of grammatical gender in both CzEng and English groups across all of the face combination conditions (all p's < .0001), indicating that participants' responses were congruent with the grammatical gender of the referent, that is when responding to an item depicting a grammatically feminine referent, participants would be more likely to choose the more feminine (or less masculine) face. There were also no significant differences between the CzEng and English groups when comparing within grammatical genders (all p's > .05), that is CzEng and English groups responded in the same way for both grammatically feminine and masculine items. This demonstrates,

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

Table 4. Pairwise comparisons for items depicting people, animals, and inanimate nouns in Experiment 3. The comparisons test the effect of grammatical gender on ratings of conceptual gender for different face combinations for CzEng and English

	estimate	*t*	*p*
Experiment 2: People			
Czech			
Words	3.550	9.788	< .001
Colour	3.653	10.071	< .001
Gray	3.495	9.636	< .001
CzEng			
Words	3.181	8.769	<.001
Colour	3.567	9.834	< .001
Gray	3.257	8.980	< .001
English			
Words	3.331	9.184	< .001
Colour	3.313	9.133	< .001
Gray	3.559	9.812	< .001
Experiment 2: Animals			
Czech			
Words	1.491	7.061	< .001
Colour	1.285	6.089	< .001
Gray	1.230	5.825	< .001
CzEng			
Words	1.351	6.402	<.001
Colour	1.472	6.972	< .001
Gray	1.284	6.082	< .001
English			
Words	.444	2.102	.036
Colour	.469	2.222	.026
Gray	.475	2.250	.024
Experiment 2: Inanimate nouns			
Czech			
Words	.531	5.277	< .001
Colour	.302	2.995	.003
Gray	.282	2.804	.005
CzEng			
Words	.401	3.981	<.001
Colour	.328	3.253	.001
Gray	.279	2.768	.006
English			
Words	.096	.956	.339
Colour	.087	.862	.389
Gray	.100	.994	.320

once more, that there is no evidence that having a grammatically gendered L1 influences participants' associations with items depicting people.

5.3.2. Animals

For the analyses of items depicting animals, we observed significant effects for CzEng participants when comparing across grammatical genders for all face combinations (all p's < .01). This indicates, as was the case in Experiments 1 and 2, that CzEng participants responded to animals in a way that was congruent with the grammatical gender of the word label in Czech. However, for English participants, we also observe a similar pattern, although it is much weaker in terms of the magnitude and

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

consistency of the effect. Specifically, we observe a significant difference in the feminine-masculine and neutral-masculine face combinations (both p's < .05), but not for the feminine-neutral combination. When interpreting the model estimates, we see for all face combinations the estimate for CzEng is almost always double the estimate for English participants. This indicates that even when an effect is observed for English participants, it is much smaller in comparison to that of CzEng participants. When comparing the language groups within grammatical genders, we observed an effect between CzEng and English for grammatically feminine items in the feminine-masculine condition (b = .438, z = 3.661, p < .001) and feminine-neutral condition (b = .245, z = 2.051, p = .040), but only a marginal effect in the masculine-neutral condition (b = .226, z = 1.874, p = .061). We observed no significant differences between CzEng and English groups for grammatically masculine items across all the face combination conditions (all p's > .05).

We again interpret these results as tentative evidence in support of the LRH, but it is also worth noting that the feminine-masculine combination is where the effect appears most saliently, suggesting that the distinction between feminine and masculine faces is more salient, in comparison to when a neutral face is presented.

5.3.3. Inanimate nouns

For the analysis of the items depicting inanimate nouns, there were no significant effects for either CzEng or English when testing for the effect of grammatical gender. Likewise, there were no significant differences when comparing across languages within grammatical genders. This indicates that unlike in Experiments 1 and 2, both CzEng and English participants show no evidence for an effect of grammatical gender influencing the decision of the faces. Thus, we do not see any evidence that would support the LRH for these items.

6. General discussion

The overarching aim of this paper was to investigate the effects of grammatical gender on the way conceptual gender is represented across speakers of Czech in L1 and L2 English, in comparison to L1 English speakers. To do this, we present data from three experiments where our focus was to use both explicit and more implicit measures of conceptual gender, analysing the data from items where the referent is a person, animal, or inanimate.

6.1. Methodological contribution

Experiment 1 took a norming approach to quantifying conceptual gender, which has become an emerging dimension of interest in the psycholinguistics literature, with a megastudy dataset available for thousands of words in grammatically gendered languages (Preininger et al., 2022; Vankrunkelsven et al., 2024) and non-grammatically gendered languages (Scott et al., 2019). If these word-based ratings are confounded in grammatically gendered languages by the presence of gender marking, then there may be problems in understanding the effects of linguistic relativity, as participants may be using the marking to bias their responses towards the grammatical gender. However, the same pattern of results was observed in the Czech and the CzEng data, suggesting that when this confound is minimized by

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

collecting ratings from translation equivalents in a non-grammatically gendered language with L2 participants, then the impact of marking may not be contributing to the ratings substantially. This is further demonstrated by the results from Experiment 2, where we observed no significant differences in the patterns of results between Czech and CzEng, but this time with words and picture stimuli. This is the first time such a large-scale norming approach has been adopted in terms of participants, languages and stimuli that focus on conceptual gender and has provided novel datasets that we hope can be reused for further empirical work on how gender is represented.

However, one issue when using a norming approach to investigate questions related to the LRH is the salient role of language and gender when making explicit conceptual gender ratings. Although our design attempted to minimize the saliency of gender by asking participants to rate other semantic dimensions, and language by using images and not just words, there still needs to be a greater effort to restrict the roles of language and gender to gain more valid insights into how grammatical and conceptual gender interact (Ramos & Roberson, 2011; Samuel et al., 2019). Thus, in Experiment 3, we used an image-face decision experiment inspired by Sato and Athanasopoulos (2018), where participants made a speeded decision between feminine/neutral/masculine faces based on an image prime. To check whether participants found grammatical gender as task-relevant, we asked participants in a post-experiment questionnaire what they thought the purpose of the experiment was. Out of all the participants, 75.58% stated stereotypes and 4.62% grammatical gender[3], suggesting that the linguistic aspect of the experiment was not the primary focus. For those items that are stereotyped for gender, this stereotyping was likely the more dominant driver of the conceptual gender association, e.g., the image of an axe (Czech: sekera) was always strongly associated with masculinity, despite it being grammatically feminine in Czech. Our approach also included face combinations that were not just simply contrasting feminine and masculine, but instead we were able to contrast neutral faces too. Although we normed the faces to show they aligned to either feminine/neutral/masculine associative gender, a categorization of gender might still have been made along the binary (see van Berlekom et al., 2024), which could have influenced the responses.

6.2. Empirical contribution

Across the three experiments, our results showed the following key patterns:

People: We observed a congruency effect, whereby participants would associate the conceptual gender with the Czech grammatical gender, irrespective of the participant's L1. There were no differences between languages in terms of the magnitude of these effects, with Czech, CzEng and English data showing no differences in any of the experiments.

Animals: The grammatical gender assigned to animals is not typically a transparent assignment, it is therefore difficult to know the biological sex of most animals

[3]The other options were: emotional expressions (7.59%), face attractiveness (4.46%), language learning (0.5%), no idea (4.29%), something else (2.97%).

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

without any prior knowledge of any subtle dimorphic differences that exist. Across all three experiments, we observed that the participants in the CzEng condition would have a congruent relationship between the grammatical gender in Czech and their ratings/decisions for conceptual gender. When comparing this to the English data, we do not see a strong congruent relationship as there is no available prior knowledge of a grammatical gender system. Additionally, we observed a pattern where grammatically feminine items were rated as significantly more feminine by CzEng participants when compared to the English participants, but we did not observe this for grammatically masculine items. Understanding this asymmetry and explaining it sufficiently will need to be addressed in future research. We would speculate that there could be a systematic relationship between the grammatical and conceptual gender of some animals in Czech that is also present as an implicit bias in English-speaking participants too, for example, a rhino is conceptually masculine for both CzEng and English participants and is grammatically masculine in Czech.

Inanimate nouns: There has been a lot of focus on inanimate nouns when investigating the LRH, given that they should have no clear biological sex and thus the relationship between grammatical and conceptual gender should provide a valid way to test the hypothesis. In our explicit rating experiments, our analyses indicated that Czech and CzEng participants did have a congruent relationship between grammatical and conceptual gender. In Experiment 1 we also observed this pattern for English participants, although it was much smaller in terms of size when compared to the Czech and CzEng effects, whereas in Experiment 2 there was no clear indication of the effect in English participants. When we assessed whether the ratings differed significantly between languages for grammatically feminine and masculine items, we only observed an effect in Experiment 1 when comparing Czech to English. In Experiment 3 we also did not observe any strong evidence for congruency in the data in CzEng or English, nor were there any language differences. Taken together, we would interpret these results as not overwhelmingly supportive of the LRH, at least for inanimate nouns.

6.3. Future directions

Although our macro-level approach allowed us to collect data on a wide range of stimuli, a more nuanced analysis would look into whether some items had a systematic relationship between grammatical and conceptual gender, whereby there is an underlying reason for the grammatical gender assignment that has been conserved over time and is resistant to change. This could explain why the English language group showed effects that aligned with Czech grammatical gender, e.g. grammatically masculine animals. One potential way to assess this would be to take a macro cross-linguistic approach, where one would test whether there is a consistency in the assignment of grammatical gender for animals (or any other nouns) across unrelated, and even related, languages. There is a growing interest in whether such 'universalities' exist (e.g. Dubenko, 2022; Williams et al., 2019, 2021), with results suggesting that (for inanimate nouns) there is evidence to support the idea that a non-arbitrary relationship can be observed, challenging the widely held assumption that assignment is arbitrary. Knowing which words in a language have a grammatical gender assignment that is correlated or uncorrelated across multiple

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

languages may provide an important methodological consideration for future research when selecting stimuli for experiments. Moreover, from a diachronic perspective, if a grammatical gender is stable or unstable over time may also be an additional factor that could reflect an underlying relationship between the grammatical and conceptual gender.

Moreover, an additional level of analysis that looks at the response times of participants when making their decisions could prove valuable. In Experiments 1 and 2 this would not be feasible as rating tasks are not typically timed, but for Experiment 3 this should be viable. Similarly, we would also be interested in assessing whether proficiency in the participant's non-grammatically gendered L2 may explain some of the variance in our data, with those participants who have lower proficiency potentially relying more on access to the grammatical gender of their L1, which may lead to stronger linguistic relativity effects. As both of these analyses would be related to more specific cognitive mechanisms of processing, we aim to address them in future work where sufficient theoretical and empirical attention can be given.

7. Conclusion

Across three experiments, we employed different paradigms that aimed to investigate how grammatical gender may influence conceptual gender. Our results highlight that there is support for the LRH when the items are animals, where we observed L1 Czech speakers congruently associating conceptual gender with the Czech grammatical gender of those animals in their L1 and non-grammatically gendered L2. However, we did not see the same clear pattern of results for inanimate nouns, suggesting that a strong linguistic relativity account is not supported by our data. The distinction between animals and inanimate nouns is critical – animals are animate, but for most, it is unclear what the biological sex of the animal is. It may be that the grammatical gender of such animals acts as an important cue to the way they are conceptually represented (Kousta et al., 2008; Imai et al., 2014; Saalbach et al., 2012; Vigliocco et al., 2005), whereas for inanimate nouns, where there is normally no information about biological sex, the effects of grammatical gender are substantially weakened or removed altogether.

Data availability statement. The data from the study, as well as the code used to produce the experiments and analyses, can be found on the Open Science Framework at https://osf.io/uky6q.

Acknowledgements. The authors would also like to thank Jan Chromý and members of the ERCEL research group at Charles University for helpful feedback as this work developed. We would also like to thank the two anonymous reviewers for their comments during the review process.

Funding statement. This research was supported by a PRIMUS grant (PRIMUS/21/HUM/015) from Charles University awarded to JB and a GAČR (Grantová Agentura České Republiky) Standard grant (23-06796S) awarded to JB. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval. This research was conducted with ethical approval obtained from the ethics committee at Charles University.

References

Almutrafi, F. (2015). *Language and cognition: Effects of grammatical gender on the categorisation of objects* (Unpublished doctoral thesis). Newcastle University, Newcastle upon Tyne, UK.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using LME4. *Journal of Statistical Software*, 67 (1), 1–48.

Belacchi, C., & Cubelli, R. (2012). Implicit knowledge of grammatical gender in preschool children. *Journal of Psycholinguistic Research*, 41, 295–310.

Boutonnet, B., Athanasopoulos, P., & Thierry, G. (2012). Unconscious effects of grammatical gender during object categorisation. *Brain Research*, 1479, 72–79.

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal. *Journal of Statistical Software*, 35.

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

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85), 5351.

Dubenko, E. (2022). Across-language masculinity of oceans and femininity of guitars: Exploring grammatical gender universalities. *Frontiers in Psychology*, 13, 1009966.

Duñabeitia, J. A., Baciero, A., Antoniou, K., Antoniou, M., Ataman, E., Baus, C., … & Pliatsikas, C. (2022). The multilingual picture database. *Scientific Data*, 9(1), 431.

Imai, M., Schalk, L., Saalbach, H., & Okada, H. (2014). All giraffes have female-specific properties: Influence of grammatical gender on deductive reasoning about sex-specific properties in German speakers. *Cognitive Science*, 38(3), 514–536.

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22, 519–534.

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2008). Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 843.

Kramer, R. (2020). Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6, 45–66.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., … & Škrabal, M. (2021). Corpus SYN, version 9 from 5. 12. 2021. Ústav Českého národního korpusu FF UK, Praha 2021. https://www.korpus.cz

Lange, K., Kühn, S., & Filevich, E. (2015). Just another tool for online studies (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, 10(6), e0130834.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Lewis, M., Cooper Borkenhagen, M., Converse, E., Lupyan, G., & Seidenberg, M. S. (2022). What might books be teaching young children about gender?. *Psychological Science*, 33(1), 33–47.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology*, 79, 328–348.

Lucy, J. A. (2016). Recent advances in the study of linguistic relativity in historical context: A critical assessment. *Language Learning*, 66(3), 487–515.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, 54.

Preininger, M., Brand, J., & Kříž, A. (2022). Quantifying the socio-semantic representations of words. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44, No. 44).

Preininger, M., Brand, J., Kříž, A. & Cehánková, M 2024 . SocioLex-CZ: Normative estimates for socio-semantic dimensions of meaning for 2,999 words and 1,000 images.

R Core Team (2023). _R: A language and environment for statistical computing_. R Foundation for Statistical Computing. https://www.R-project.org/.

Ramos, S., & Roberson, D. (2011). What constrains grammatical gender effects on semantic judgements? Evidence from Portuguese. *Journal of Cognitive Psychology*, 23(1), 102–111.

Saalbach, H., Imai, M., & Schalk, L. (2012). Grammatical gender and inferences about biological properties in German-speaking children. *Cognitive Science*, 36(7), 1251–1267.

Samuel, S., Cole, G., & Eacott, M. J. (2019). Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review*, 26(6), 1767–1786.

https://doi.org/10.1017/langcog.2025.3 Published online by Cambridge University Press

Sato, S., & Athanasopoulos, P. (2018). Grammatical gender affects gender perception: Evidence for the structural-feedback hypothesis. *Cognition*, 176, 220–231.

Sato, S., Casaponsa, A., & Athanasopoulos, P. (2020). Flexing gender perception: Brain potentials reveal the cognitive permeability of grammatical information. *Cognitive Science*, 44(9), e12884.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51, 1258–1270.

Sera, M. D., Elieff, C., Forbes, J., Burch, M. C., Rodríguez, W., & Dubois, D. P. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General*, 131(3), 377.

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2023). Rating norms should be calculated from cumulative link mixed effects models. *Behavior Research Methods*, 55(5), 2175–2196.

Thierry, G. (2016). Neurolinguistic relativity: How language flexes human perception and cognition. *Language Learning*, 66(3), 690–713.

van Berlekom, E., Sczesny, S., & Sendén, M. G. (2024). Toward visibility: Using the Swedish gender-inclusive pronoun hen increases gender categorization of androgynous faces as nonbinary. *Journal of Language and Social Psychology*, 43(5–6), 525–543.

Vankrunkelsven, H., Yang, Y., Brysbaert, M., De Deyne, S., & Storms, G. (2024). Semantic gender: Norms for 24,000 Dutch words and its role in word meaning. *Behavior Research Methods*, 56(1), 113–125.

Vigliocco, G., Vinson, D. P., Paganelli, F., & Dworzynski, K. (2005). Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General*, 134(4), 501.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT Press.

Williams, A., Cotterell, R., Wolf-Sonkin, L., Blasi, D., & Wallach, H. (2019). Quantifying the semantic core of gender systems. arXiv preprint arXiv:1910.13497.

Williams, A., Cotterell, R., Wolf-Sonkin, L., Blasi, D., & Wallach, H. (2021). On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9, 139–159.

Cite this article: Brand, J., Preininger, M., Kříž, A., & Cehákova, M. (2025). Feminine fox, not so feminine box: constraints on linguistic relativity effects for grammatical and conceptual gender, *Language and Cognition*, 17, e34, 1–26. https://doi.org/10.1017/langcog.2025.3