SocioLex-CZ: Normative estimates for socio-semantic dimensions of meaning for 2,999 words and 1,000 images

Abstract

Words live social lives. When we encounter words, we not only activate the social information provided by the speaker, but also the rich semantics of the word's meaning, and quantifying this information is a key challenge for the cognitive and behavioural sciences. Although there are many resources available that quantify affective and sensorimotor information, there are relatively fewer resources available that provide information on social dimensions of meaning. We present the SocioLex-CZ norms, where the primary focus is on socio-semantic dimensions of meaning. Across two experiments, we introduce normative estimates along 5 dimensions – gender, political alignment, location, valence and age – for a large set of Czech words (Experiment 1) and images (Experiment 2) from 1,709 participants. We provide a series of analyses that demonstrate the norms have good reliability, as well as presenting exploratory analyses of how the variables interact with one another within and between words/images. These norms present the first dataset that looks at socio-semantic representations at scale, which we hope will be used for a range of novel and multidisciplinary applications and subsequently opening up new pathways for innovative research. We make the data, code and analysis available at https://osf.io/pv9md/?view_only=ccbd13b7afb141628b51e0d0a83f9037 and also provide an interactive web app at https://tinyurl.com/sociolex-cz-app.

Keywords: Semantics; Norms; Socio-semantics; Concepts; Representation

OSF repository:
https://osf.io/pv9md/?view_only=ccbd13b7afb141628b51e0d0a83f9037

Total word count: 9,793
Abstract word count: 202

1. Introduction

Words have rich and diverse meanings that vary along different semantic dimensions, some of which contribute substantially to their semantic representation of certain concepts, whilst others contribute less so. For example, the sensory dimension of olfaction is very important for the word aroma, but is less important for amoeba (Lynott et al., 2020). Quantifying how meaning can be represented has been a crucial aim of cognitive science, psycholinguistics, and other interdisciplinary sciences for decades, with the pioneering work of Osgood, Suci and Tannenbaum (1957) popularising the practical applications of collecting datasets of subjective ratings (or norms) along definable dimensions of meaning for lists of words.

Recent interest in such datasets has been exemplified through the growing body of studies publishing rating norms for a wide range of dimensions that capture specific aspects of meaning (e.g. humour: Engelthaler & Hills, 2018; iconicity: Winter et al., 2024; survival: Alonso, Díez & Fernandez, 2021), even investigating different language settings (e.g. multiple languages: Łuniewska et al. 2016, 2019; sign languages: Sehyr et al., 2021; second languages: Ferré et al., 2022; Imbault et al., 2021). Despite this diversity in the literature, there has traditionally been a concentrated empirical focus on two clusters of semantic dimensions, specifically affective (e.g. valence, arousal and dominance: Warriner et al., 2013), and sensorimotor (e.g. body-object interaction: Pexman et al., 2019; concreteness: Brysbaert et al., 2014; perceptual and action: Lynott et al., 2020), with these norms facilitating research on how languages are learnt, processed, and represented in the brain. However, there is growing interest to move beyond psycholinguistic, affective and sensorimotor dimensions of meaning, and instead look towards other theoretically important and empirically valuable properties of meaning (see e.g. Binder et al., 2016), to further inform models of semantic representation and open up new avenues for researchers to better understand linguistic and cognitive processes.

One emerging perspective posits that social experiences, contexts and interactions play an important role in explaining how certain concepts can be grounded in the human conceptual system, with a particular emphasis placed on the grounding of abstract concepts (Barsalou, 2020; Borghi et al., 2019; Pexman, Diveica & Binney, 2023). To date, there are two large-scale studies that have quantified the socialness of thousands of words in English (Diveica, Pexman & Binney, 2023) and Chinese (Wang et al., 2023), whereby socialness is defined broadly to capture a range of potential features that are socially relevant, e.g. "*a social characteristic of a person or group of people, a social behaviour or interaction, a social role, a social space, a social institution or system, a social value or ideology, or any other socially relevant concept*" (Diveica et al., 2023, p. 463), and has been shown to have a facilitatory effect in a range of lexical processing tasks (Diveica et al., 2024).

Although the expansive definition provided for the socialness norms facilitates the study of a complex construction through a single quantitative variable, it may present an issue for researchers who are looking to target a specific aspect of social experience, context or interaction. For example, in the Diveica et al. (2023) English socialness norms, some of the words with the highest ratings are *romance, socialism, mother, festival* and *funeral*, highlighting the breadth of representations that socialness as a variable can capture, whilst also revealing the acute differences that exist between the concepts in terms of the specific types of social information that is encoded – *festival* and *funeral* are both social events, but the semantics are normally quite contrastive in terms of the social information they represent. Thus, an alternative approach has been to investigate specific attributes that are socially relevant and can be theoretically defined, providing a more nuanced quantification of socially pertinent dimensions of representations.

One example of a socially relevant dimension being examined in this way is semantic gender, where participants are asked to rate the extent to which a word's meaning is associated to feminine, masculine or neutral attributes. Semantic gender was included as a scale in Osgood et al.'s (1957) original semantic differential work, which only focused on a very small set of words (see Vankrunkelsven et al., 2024 for a more detailed overview of other studies that have normed semantic gender). However, with the renewed attention given to collecting mega-study datasets, there are now conceptual/semantic gender ratings available for 5,500 (Scott et al., 2019) and 2,373 (Lewis et al., 2022) English words, as well as 24,000 Dutch words (Vankrunkelsven et al., 2024). Despite these datasets still being relatively new additions to the larger catalogue of norms available to researchers, they have provided a critical bridge between the social and cognitive sciences, facilitating novel investigations into multidisciplinary domains of research. For instance, by uncovering sources of gender biases present in children's literature (Lewis et al., 2022) or language more generally (Lewis & Lupyan, 2020), in addition to testing effects of linguistic relativity, whereby grammatical gender may influence the way that semantic gender is represented (Brand et al., under review; Vankrunkelsven et al., 2024).

Indeed, the field of sociolinguistics has long been demonstrating the importance of socially meaningful variables that can be used to characterise groups of individuals, (e.g. gender, age, location, political alignment) to study language use, variation and change (see e.g. Labov, 2011; Trudgill, 1974). Whilst such variables are the focus for studies at the population level (i.e. which words or variants are produced by females/males or younger/older age groups), there has been relatively little work looking at how these factors operate at the semantic level (i.e. which words have meanings that are associated with femininity/masculinity or younger/older age groups). Interestingly, when studies have used stimuli that have been chosen because of their associations to socially relevant dimensions of meaning, the findings have helped to inform models of language and memory. For example, Walker and Hay (2011) demonstrated that lexical access to words that are socially skewed towards either younger or older individuals (e.g. *internet* and *libraries*) is faster when the word is said by an age-congruent voice, with these results extending to associations for gender and non-spoken stimuli, such as images (Hay et al., 2019).

Taken together, there is growing evidence that semantic variables that are linked to socially meaningful attributes – or socio-semantics – are important features of the human conceptual system. Although there has been attention given to the dimension of semantic gender, this is only one possible source of socio-semantic information, and there are several other dimensions that are currently underexplored. Moreover, when there are datasets that quantify social dimensions, they tend to look exclusively at orthographically presented word forms, which substantially limits researchers who might want or need to design experiments that do not use linguistic stimuli. Therefore, if the study of socio-semantics is to accelerate then there needs to be more coverage in terms of the types of dimensions and the types of stimuli available as norms.

2. The SocioLex-CZ Norms

In this paper we introduce the SocioLex-CZ norms, the first large-scale dataset that quantifies the socio-semantic representations across four distinct dimensions: gender, location, political alignment and age. Additionally, the dataset also contains information on an affective dimension - valence, which was included as there are several existing studies with norms available and can act not only as a new dataset for Czech, but also as a control variable to facilitate our reliability analyses.

Across two experiments, we collected data for a diverse range of 2,999 words in Czech (Experiment 1), with an exploratory analysis of how the dimensions relate to each other, as well as how they compare to existing datasets in other languages. We also collected data for 1,000 images (Experiment 2), with another exploratory analysis of how the different stimuli types may modulate the ratings.

The primary motivation for developing the SocioLex-CZ dataset is to provide researchers with a novel resource that captures the semantic representations of stimuli along several theoretically meaningful social dimensions for both words and images. Moreover, this is the largest dataset to examine speakers of the Czech language, opening up new possibilities for researchers interested in studying a language with a rich morphology and a distinctive recent geo-political history.

The primary aims of the current study are:

- Introduce the specific dimensions of interest and the methodology used for data collection and data pre-processing
- Describe and validate the summarised dataset of normative ratings
- Explore the relationships between the different dimensions and the different stimuli versions (words and colour/grayscale images)
- Outline future directions and applications for the study of socio-semantics

3. Experiment 1: Word Ratings

3.1 Methods

Stimuli

The stimuli consisted exclusively of 2,999 written words (1,902 nouns, 766 adjectives and 331 verbs). The word list was constructed with the aim to include items that refer to a broad range of socially relevant and culturally salient concepts, across a range of different parts of speech. This was achieved by exploring word lists from published datasets. First, we took all 501 Czech word labels from the Multilingual Picture Database (Duñabeitia et al., 2022), these were the most dominant names for coloured drawings of everyday objects. Second, we took all 822 Czech role nouns used in Misersky et al (2014), which focussed on gender stereotypicality. Third, we took Czech translations of the Leipzig-Jakarta list (Haspelmath & Tadmor, 2009), which contains 100 basic cultural concepts. Fourth, we took 400 translation equivalents of German adjectives from Grühn and Smith (2008), which focussed on a range of semantic rating scales, including age. Fifth, we took the 300 most frequent verbs from the Czech National Corpus SYN2020 (Jelínek et al., 2021), which consists of contemporary written Czech. Finally, we included the remaining 876 items during a brainstorming session between the researchers.

Since Czech is a grammatically gendered language, both masculine and feminine variants were included whenever possible. This concerned mostly role-nouns (e.g. *pekař/pekařka* refer to male/female baker), nouns referring to animals (e.g. *medvěd/medvědice* refer to male/female bear) and adjectives (e.g. dřevěný/*dřevěná* are used for grammatically masculine/feminine heads in a phrase). Nouns and adjectives were presented in nominative singular, whereas verbs were in their infinitive form.

The list of words was pseudo-randomly divided into 100-word subsets. Each subset contained approximately the same number of nouns, adjectives and verbs. We also controlled the distribution of grammatically gendered words by ensuring no subset contained both grammatically feminine/masculine variants of the same word, with the number of grammatically feminine/masculine adjectives and nouns being roughly comparable across lists. All lists were further complemented with four phonotactically plausible pseudowords (e.g. *tontota*), which served as control words for an attention check, as well as a calibrator word for each socio-semantic dimension, which was always the first word presented in the list. The calibrator words were chosen on the basis of data from a pilot experiment, whereby participants reliably rated the words towards a specific direction on the scale. The calibrator words were: GENDER - *náhrdelník* (necklace), LOCATION – *metro* (subway/underground train), POLITICAL – *tradice* (tradition), VALENCE – *šikanovat* (to bully), AGE – *důchod* (pension).

Participants

An initial sample of 1,475 participants completed the experiment. 1,275 participants were recruited from a university-wide student participant pool at Charles University in the Czech Republic, with all students receiving course credit for taking part. Ethical approval was given by Charles University. From this sample, we excluded 48 participants who reported that their native language was not Czech. An additional

200 participants were also recruited using Prolific (www.prolific.com), who were paid £5.00 for taking part and declared that they were currently a university student and had Czech as their L1. We also decided to only retain participants who reported their age to be between 18-30 years old, which meant excluding 128 participants altogether. The motivation for restricting the age was to ensure our dataset would be comparable to other existing datasets with predominantly young adults (e.g. Vankrunkelsven et al., 2024). After a series of data quality checks (see Data Cleaning section below), we excluded a further 45 participants, resulting in a final sample of 1,254 participants (936 females, 312 males, 6 non-binary/self-report, median age = 21, *SD* = 2.22, range = 18–30).

All participants completed a demographic questionnaire before starting the experiment. This was designed to collect a more detailed demographic profile of the participants and included standard questions, e.g. age, categorical gender (female/male/non-binary/self-report), highest educational qualification and native language, but also questions about their perceived socio-demographic profile. This included 7-point Likert scales where participants self-assessed their gender stereotypicality (typical male - typical female), character (very optimistic - very pessimistic), location affiliation (very urban - very rural), and political alignment (very liberal - very conservative). All scales were bipolar with a neutral midpoint. This data is visualised in Figure 1. We include these variables in the supplementary materials, but note the aims of this paper are not focussed on providing a detailed analysis of how demographic differences may influence ratings. We return to this in the General Discussion and address this in ongoing work where sufficient depth and attention can be given.

Procedure

The experiment was designed and distributed online using Qualtrics (https://www.qualtrics.com). Participants first read a detailed instruction page, provided informed consent and completed the demographic questionnaire described above. The original Czech and translated versions are available in the supplementary materials. Participants were then asked to rate each of the words from a randomly selected list (each containing 100 words, 1 calibrator word and 4 pseudowords). Specifically, participants were asked to rate the words based on how they associated the meaning of the word according to each of the following dimensions:

- GENDER (very masculine - very feminine. Czech version: GENDER *silně mužské - silně ženské*)
- LOCATION (very rural - very urban. Czech version: PROSTŘEDÍ *velmi venkovské - velmi městské*)
- POLITICAL (very liberal - very conservative. Czech version: POLITICKÉ PŘESVĚDČENÍ *velmi liberální – velmi konzervativní)*
- VALENCE (very positive - very negative. Czech version: EMOCE *velmi pozitivní - velmi negativní*)
- AGE (divided into categories of 0-6, 7-17, 18-30, 31-50, 51-65, 66-80, and 81+ years. Czech version: VĚK)

Participants were presented with one dimension at a time, which contained all the words from one of the subsets. The order of presentation for dimensions and words was randomised for each participant (apart from the calibrator word, which was always presented first, and the age dimension, which was always presented as the last dimension). All dimensions (apart from AGE) were rated using 7-point Likert scales, each with a neutral midpoint, see Figure 2A. For the age dimension, participants were instructed to select as many age categories as they felt applicable using checkboxes, meaning multiple categories or no categories could be chosen. This was done so that participants could provide a more detailed age selection that was not restricted to just one forced choice, e.g. the word *pensioner* would likely activate both 66-80 and 81+ categories, whereas *paper* would activate no categories. All dimensions had the option to skip a word if the meaning was not known (toto slovo neznám)

3.2 Data Cleaning

All cleaning, processing and analyses reported below were conducted using R version 4.3.1 (R Core Team, 2023). All the code and data (raw and processed) can be found in the supplementary materials.

We carried out a number of data quality checks to ensure that we detected any low-quality responses/participants, this process is visualised in Figure 3. The code used to run these checks can be found in the supplementary materials.

1. *Word knowledge:* We checked for any participants who reported not knowing over 20% of the words they were presented with, however there were no such participants.
2. *Straightlining:* We inspected the variance in the participant's responses to identify straightliners, i.e. participants who selected the same response for more than 80% of the items (Kim et al., 2019; Winter et al., 2024). If a participant straightlined on more than 3 out of the 5 dimensions we decided not to include them in the final dataset, resulting in the exclusion of 17 participants. Additionally, we removed a participant's response to words on a specific dimension if they straightlined for over 95% of the words in this dimension, resulting in the exclusion of 31 responses.
3. *Calibrator words:* We analysed the responses to the calibrator words for each of the 5 dimensions to ensure that participants could demonstrate that they were using the rating scales as expected. For each calibrator word we identified participants who responded to the word in the opposite side as expected, e.g. the word *metro* for the location dimension was expected to have a response of urban, so if a participant responded to this word as rural in any way (i.e. slightly rural, rural or very rural) this would be flagged. Any participant who responded to 2 or more calibrator words in the opposite direction to our original expectations would be excluded, resulting in the exclusion of 10 participants.

4. *Control words:* We also analysed the responses to the control words (pseudowords that were phonotactically legal in Czech), excluding any participant who consistently rated all four of these words throughout the experiment, instead of choosing the '*I don't know this word*' option. This resulted in the exclusion of 15 participants.
5. *Timing:* We inspected the time it took for each participant to complete the ratings along each of the dimensions. We first transformed the timings to natural log values, then calculated the mean and standard deviations for each of the dimensions. A participant was then excluded completely if they were quicker than the mean by 2.5 times the standard deviation for 2 or more dimensions (this was always less than 116 seconds, with the mean values ranging between 349-451 seconds). This resulted in the exclusion of 3 participants. Additionally, we removed a participant's response to words on a specific dimension if their timing was quicker than the mean by 2.5 standard deviations, but retained all their other responses, resulting in the removal of data for 13 responses to individual dimensions. There was no upper cut-off point, since no time constraints were mentioned in the instructions and participants could take however long they needed to finish the experiment.

3.3 Data Summaries

We processed the raw data in a number of different ways to generate the final summary data of the SocioLex-CZ norms. We processed the dimensions of GENDER, LOCATION, POLITICAL and VALENCE differently to the AGE dimension as the scales differed.

3.3.1 GENDER, LOCATION, POLITICAL and VALENCE

Descriptive statistics

For the dimensions of GENDER, LOCATION, POLITICAL and VALENCE the values were first transformed to numeric scales, ranging from -3 (very masculine/rural/liberal/negative) to 3 (very feminine/urban/conservative/positive), with 0 being the neutral midpoint. Responses where the participant responded that they did not know the word were coded as NA values. From this data we calculated the mean, SD, number of responses overall, number of responses where the word was known and the proportion of known responses, for each of the words across the dimensions. See Figure 4 for a visualisation of the data and Table 1 for sample size summaries. We can see that the majority of ratings are around the neutral midpoint, with a skew towards more positive ratings for the VALENCE dimension, which has also been reported in other studies (e.g. Warriner et al., 2013). Table 2 gives examples of the items with the most extreme values for each dimension.

Proportions and entropy

We also provide additional descriptive statistics for each of the items that may be of interest to researchers. Specifically, raw counts are available for each of the items for each of the dimension's scale points, allowing for a more nuanced description of how often each response was selected, which have also been transformed into proportional values. Additionally, we used the count data to calculate the Shannon entropy (Shannon, 1948) for each item, along each dimension, providing an alternative measurement to SD. Entropy was included because it captures uncertainty in the ordinal Likert scale, whereas SD is better suited to measuring variance in numeric scales.

Latent means

When aggregating rating data from Likert scales, many datasets are published with the means for each item, where the ratings are treated as integer values, e.g. very negative on our valence scale as -3. However, this approach may not be the most suitable as means do not preserve the ordinal nature of the Likert scales being used (see Liddell & Kruschke, 2018; Veríssimo, 2021). Therefore, we followed the guidance of Taylor et al. (2023) and modelled the participant responses using Cumulative Link Mixed-effects Models with the `ordinal` package (Christensen, 2018) in R. These models account for variation that is introduced from the participant response biases, and thus provide a more accurate estimation of a normative estimate for each item. We modelled the ratings for each of the dimensions with a separate model, predicting the participant responses (coded as an ordinal factor, i.e. -3 < -2 < -1 < 0 < 1 < 2 < 3) with random intercepts for item and participant[1]. From this we were able to extract the random intercepts for each of the items, providing us with a numeric estimate of the latent mean. We inspected the correlation between the latent means and the standard means, with the two variables correlating almost perfectly (all *r*'s > .99).

3.3.2 AGE

To obtain a normative estimate from the age ratings (which were not collected using Likert scales, but instead categorical checkboxes where a single, multiple or no option could be selected), we needed to apply a number of processing steps to generate the summary values.

Weighted proportions

We transformed the raw categorical responses to calculate a weighted proportional value for each of the age categories. This was based on the number of age categories selected by a participant for each item (1/n_selected_categories), i.e. if a participant selected 2 age categories for an item, each selected category would have a weighted rating of 0.5 and all others would have 0, if 1 category was selected then the value would be 1, if all 7 categories were selected the value would be 1/7, if no categories were selected then an additional 'no age' category would have 1. From the weighted values we could then calculate a weighted proportion for each of the resulting 8 age categories within each item, by summing the weighted proportions and dividing that value by the number of participants who rated the item (excluding responses where the participant did not know the word). This gave us a value between 0 (no responses for the category) and 1 (all participants only selected the one category), which can be

[1] Simplified R syntax for latent mean models: `clmm(rating_ordinal ~ 1 + (1|item) + (1|participant), data = dimension_ratings, link = "probit")`

used as an estimate of the associative strength for an item in each of the age categories.

Mode category

Based on the proportion values, we were also able to calculate the mode age category for each item, i.e. the category that had largest proportional value, which can be used as a categorical summary variable that captures the most likely age association, e.g. *školka* [kindergarten] has an age mode of 0-6. If there were multiple mode categories for an item, the category was randomly selected from the possible categories, e.g. *mechanička* [female mechanic] had mode values of 18-30 and 31-50, but was assigned 18-30. Since we also took into account instances where no categories were selected, the mode could take the form of 'no age'.

Principal Component Analysis (PCA)

To create a summary value for the AGE dimension, we used the weighted proportion values (where each row is an item, and each column is an age category) as a 2,999 x 8 multidimensional space. This acted as the input data for a PCA, which allows us to reduce down the number of dimensions in the space by identifying underlying structure in the data. The result of the PCA is a set of Principal Components (PCs) that can be interpreted based on how the original variables (age categories) are loaded, i.e. which variables co-vary together within each PC. Additionally, each item will have a Principal Component Score that reflects the strength of association to each of the components, which will act as the individual item norm for each of the PCs (see Venables & Ripley, 2013; Brand et al., 2021; Wilson Black et al., 2023). The results of the PCA provided us with three main principal components (PC), each of which explained over 15% of the variance in the data and 74.19% of the overall variance. We will interpret the PCs individually as PC1, PC2 and PC3. The visualisations that can be used to guide the interpretation of the PCA results are in Figure 5.

PC1: younger/older – This PC accounted for 33.82% of the total variance and is interpreted as capturing age categories that are under the age of 30 (positive PC1 scores) or over the age of 30 (negative PC1 scores). We highlight that this is not a linear relationship, the most extreme PC1 scores are not directly related to the youngest or oldest age categories, but instead represents a more general distinction between young and old associations, e.g. *beďar* [pimple] has the highest PC1 score of 3.79, whereas *penze* [pension] has the lowest score of -5.68.

PC2: Middle age – This PC accounted for 24.82% of the total variance and is interpreted as capturing age categories that are related to middle age and those which are not. This appears to be a U-shaped relationship, where negative values are most association with the 31-50 age category and more positive values with the youngest (0-6) and oldest age categories (66-80 and 81+), e.g. *statistik* [male statistician] has a PC2 score of -3.72, whereas *hřbitov* [graveyard] has 4.86.

PC3: No age – This PC accounted for 15.54% of the total variance and is interpreted as capturing items that have no age associations and those that do. The negative PC3 scores represent stronger associations to no age, whereas positive scores appear to represent stronger associations to any age category, but specifically the oldest age categories e.g. *přehrada* [dam] has a PC3 score of -3.31, whereas *senior* [male senior citizen) has 4.34. We note that the individual weighted proportion scores for the no age category would likely serve as a more accurate representation of this PC, as it can be interpreted more straightforwardly, whereas the PCA introduces more complexity when interpreting the positive values, see Figure 5B.

3.4 Analysis and Results

Reliability analysis

The internal consistency of the individual rating scales was assessed using Cronbach's alpha (Cronbach, 1951), allowing us to measure the reliability of our items and indicate how well they measure the same underlying constructs – in this case, each of the socio-semantic dimensions. We calculated alpha values for each list of words[2] individually for the GENDER, LOCATION, POLITICAL and VALENCE dimensions, where only one option per word could be selected. In the case of the AGE dimension, participants could choose as many options as they liked which yielded a large variance in individual ratings, rendering the calculation of Cronbach's alpha unsuitable. The range of alpha values across the lists were: GENDER = [.92, .95], LOCATION = [.89, .95], POLITICAL = [.88, .95], VALENCE = [.90, .95]. This indicates that the rating scales can be considered to have high reliability.

Correlations between variables

We ran an exploratory correlational analysis to inspect whether the mean values for each item are correlated across the different socio-semantic dimensions. As the means across all the dimensions had a reasonably normal distribution, we conducted Pearson's correlations between each of the dimensions, see Figure 6 for the visualisation of the results. Note that as this was an exploratory analysis and we were not directly testing any research hypotheses, we purposefully do not report p values from the tests and will instead simply describe patterns in the data where the correlations were strongest, i.e. r > |.18|.

The strongest correlation was between POLITICAL and AGE PC_1 (r = -.533), indicating that liberal items were more likely to be associated with youth, whilst more conservative items with older age. POLITICAL was also related to LOCATION (r = -.354: more liberal items are associated with urban environments, conservative with rural), VALENCE (r = -.317: more liberal items are associated to positive valence, conservative with negative), AGE PC_2 (r = .253: more liberal items are associated with middle age) and GENDER (r = -.185: more liberal items are associated with femininity, conservative with masculinity). LOCATION was also correlated with AGE PC_2 (r = -.468), indicating that

[2] We excluded two words (*talmud* [Talmud] and *imám* [Imam]) from their respective lists, due to these items being known by < 50% of the participants, resulting in missing values. We also excluded the control and calibrator words from this analysis, so that we only analysed the items that were assigned to the specific lists.

more urban items were associated with middle age. GENDER and VALENCE were also correlated (r = .183), with more feminine items associated with positive ratings, whereas masculinity was associated with negative ratings, which was also reported for Scott et al's (2019) English data (who collected both gender and valence ratings) and Vankrunkelsven et al.'s (2024) and Moors et al.'s (2012) Dutch data (based on their gender ratings and Dutch valence ratings from Moors et al (2012)).

We would like to highlight that not all of these relationships were linear, indeed in Figure 6 we fitted a loess smooth to the data to demonstrate this more nuanced interpretation of how the dimensions may relate to one another, especially at the more extreme ends of the distributions. However, a more detailed examination of these relationships is outside the scope of the current paper.

Correlations with other languages

As an additional level of analyses, we were interested in looking at how our data correlates to other existing datasets with the same underlying dimensions, testing the hypothesis that our data will correlate with other datasets in different languages. As this is the first large-scale norming dataset of variables for Czech words, we unfortunately could only assess whether our dataset correlates to existing datasets in other languages, instead of other datasets in Czech. We included the valence dimension in our norms, as the valence norms for English words by Warriner et al (2013) have been used extensively by researchers since they were published, and offered a good baseline for comparison. Additionally, Scott et al's (2019) dataset includes both valence and gender norms for English words, which provided us with an additional resource that was published more recently.

Subsequently, we coded all 2,999 Czech words for translation equivalents in English (which were translated and checked by two native Czech speakers, both highly proficient in English), providing us with a basis to compare across the different datasets. We then filtered the data so that we had the original Czech mean ratings for valence and the mean ratings of the translation equivalents in English. Using this dataset, we ran Pearson's correlations to inspect the relationship between the Czech and the English ratings. From Warriner et al's (2013) valence ratings, we had 1,546 comparable items and there was a very strong correlation, $r(1544) = .873$, $p < .001$. We took both the valence and gender ratings from Scott et al (2019) with 1,166 comparable items and again found a very strong correlation for valence, $r(1164) = .916$, $p < .001$ and a strong correlation for gender, $r(1164) = -.658$, $p < .001$[3].

4. Experiment 2: Image ratings

Typically, semantic norms are only published for items that are represented orthographically, i.e. as words. However, we wanted to also provide socio-semantic norms for image stimuli, to allow researchers to utilise the resource in a wider range of experimental settings, e.g. tasks that might want to avoid text-based stimuli. Thus, we also provide an additional dataset of socio-semantic norms relating directly to image stimuli in colour and grayscale.

4.1 Methods

Stimuli

Our stimuli consisted exclusively of 1,000 images taken from the Multilingual Picture Database (Duñabeitia et al., 2022). This database contains drawings for commonly encountered nouns from a range of semantic categories (e.g. role nouns, animals, tools, clothing), with naming data from a number of different languages, including Czech. The Multilingual Picture Database contains 500 unique images, but we decided to include both colour and grayscale versions of each image in our stimuli (providing us with 1,000 images in total). This decision was motivated primarily by the need for diverse stimuli sets that can be used in experimental designs that may only require colour/grayscale stimuli.

[3] Note that the correlation here is negative as the Scott et al (2019) gender ratings had masculine/feminine on opposite sides of the scale to our ratings.

Participants

An initial sample of 495 participants completed the experiment. All participants were recruited from a university-wide student participant pool at Charles University in the Czech Republic, with all students receiving course credit for taking part. From this sample, we excluded 11 participants who reported that their native language was not Czech. We again decided to only retain participants who reported their age to be between 18-30 years old, which meant excluding 19 participants. After a series of data quality checks (see Data Cleaning section below), we excluded a further 10 participants, resulting in a final sample of 455 participants (373 females, 75 males, 7 non-binary/self-report, median age = 21, *SD* = 1.95, range = 19–29). All participants completed the same demographic questionnaire as described in Experiment 1, with the data visualised in Figure 1.

Procedure

The procedure roughly matched the same procedure for Experiment 1, but with a small number of changes. Participants were presented with one image at a time and were asked to rate the image along the dimensions of GENDER, LOCATION, POLITICAL and VALENCE using the same 7-point Likert scales, see Figure 2B. The AGE dimension was not included for this experiment as it was not possible to incorporate the checkbox input required for this variable in the design using Qualtrics. Participants could also select an option (*nevím, o co se jedná [I don't know what this is]*) if they did not know what the image was. All participants saw an image of a mouse as their first item, which was selected as it was very familiar based on Duñabeitia et al's (2022) data, then rated a set of 100 items presented in a randomised order.

4.2 Data Cleaning

We carried out a number of data quality checks to ensure that we detected any low-quality responses/participants, this process is visualised in Figure 3. This was similar to the process used in Experiment 1, but we did not implement the calibrator and control item steps as there was only one calibrator item and no controls (as control items for images was not possible).

1. *Image knowledge:* We checked for any participants who reported not knowing over 20% of the images they were presented with, however there were no such participants.
2. *Straightlining:* If a participant straightlined on more than 2 out of the 4 dimensions we decided not to include them in the final dataset, resulting in the exclusion of 10 participants. Additionally, we removed a participant's response to words on a specific dimension if they straightlined for over 95% of the images, resulting in the exclusion of 1,919 responses.
3. *Timing:* We inspected the time taken to complete the whole experiment by each participant, by log transforming the durations. We again applied a filter of any participant who was quicker than the mean duration by 2.5 standard deviations, but there were no such participants. Once again, there was no upper cut-off point.

4.3 Data Summaries

We calculated the same summary statistics for each of the images for both colour and grayscale as those reported for Experiment 1. See Figure 4 for a visualisation of the data and Table 1 for sample size summaries.

4.4 Analysis and Results

Reliability analysis

The internal consistency of the individual rating scales was assessed using Cronbach's alpha (Cronbach, 1951). We calculated alpha values for each list of images[4]. The range of alpha values across the lists were: GENDER = [.91, .95], LOCATION = [.88, .92], POLITICAL

[4] We excluded five images (*PICTURE_146_gray, PICTURE_192_colour, PICTURE_192_gray, PICTURE_616_colour* and *PICTURE_616_gray*) from their respective lists, due to these items being known by < 50% of the participants, resulting in missing values.

= [.88, .93], VALENCE = [.89, .95]. This indicates that the rating scales for the images can be considered to have high reliability, as was the case in Experiment 1.

Correlations between item versions

In order to assess whether the socio-semantic norms differ across stimuli types (i.e. words/colour images/grayscale images), we inspected the correlations between the ratings for each of the dimensions. We analysed the data using Pearson's correlation coefficients, with each stimuli type comparison (i.e. colour~grayscale, colour~words, grayscale~words), with all *r*'s > .76, see Figure 6.

Colour vs Grayscale images: The strongest correlations were found between the two image versions, with all *r*'s > .87. Whilst this strong relationship might be expected given the stimuli only differ in terms of their visual appearance, there are items that exhibit notable variation. For example, in the GENDER dimension, the colour and grayscale versions of the image for a dressing gown have substantially different ratings, where the colour version displays the item in pink and this has a more feminine association (*M* = 2.02), whereas the grayscale version is much more neutral (*M* = -0.06), see item A in Figure 6. This variation might be explained by the stereotyping of pink clothing being associated more with femininity, and when the colour bias is removed, the gender association bias might also be affected. However, there appears to be other sources of variation between colour and grayscale items, for example the colour image of a cut (wound) (see item B in Figure 6) has a much more negative rating for VALENCE (*M* = -2.00) than the grayscale version (*M* = -0.58), but the grayscale version is much more liberal (*M* = -1.31) than the colour version (*M* = 0.05) in the POLITICAL dimension. This might be explained as a conceptual difference, as the grayscale image may be interpreted as a tattoo rather than a cut/wound, leading to two different semantic representations (note that the Duñabeitia et al (2022) data were only normed based on the colour versions of the images).

Images vs words: The correlations between colour/grayscale images and words were slightly weaker than the colour~grayscale relationships, but were still interpreted as being very strong, indicating that - in general – images and words with the same conceptual meanings in our dataset capture very similar socio-semantic representations. However, there is again variation in the dataset, indicating that the alignment between ratings for words and images is not always stable. For example, *župan* is the Czech word for dressing gown and was rated as weakly associated to femininity (*M* = 0.78), whereas it was much more strongly associated to femininity in the colour version (*M* = 2.02), but close to neutral in the grayscale version (*M* = -0.06). Moreover, the Czech word for cut/wound is *rána* and was rated very similarly to the colour version along all dimensions, but varied considerably to the grayscale version for VALENCE and POLITICAL dimensions, again suggesting that the grayscale version may have been interpreted as something conceptually different to the colour and word versions.

5. General Discussion

The SocioLex-CZ norms present a novel and innovative tool that can provide an important resource for a diverse range of scientific applications. The dataset is the largest known resource that captures conceptual associations along the dimensions of GENDER, LOCATION, POLITICAL, VALENCE and AGE, with a number of summary variables and the underlying raw data made freely available. Moreover, we provide ratings not only for a large set of words in Czech, which has typically not received much coverage in terms of the norms currently available, but also provide ratings for both colour and grayscale images, expanding the range of potential research applications that the norms can be used for. It is hoped that the dataset will create new synergies between the psychological, social and language sciences by opening up new research questions relating to socio-semantics, where our understanding of socially meaningful dimensions of semantic representations is explored and investigated in more detail.

Researchers have long been interested in how gender is represented semantically, with large-scale norms recently becoming available in English (Lewis et al., 2022; Scott et al., 2019) and Dutch (Vankrunkelsven et al., 2024). Our Czech norms add to this growing body of work, and as Czech is a grammatically gendered language, they may be of interest to researchers investigating whether grammatical and semantic gender interact, i.e. the Linguistic Relativity Hypothesis (see Samuel, Cole and Eacott, 2019). Moreover, the large number of items in our norms may also facilitate more work that investigates how abstract concepts are grounded in social aspects of meaning (Barsalou, 2020; Borghi et al., 2019; Pexman, Diveica & Binney, 2023), with our LOCATION, POLITICAL and AGE dimensions offering novel ways to quantify socio-semantic representations. For example, words like *demokracie* [democracy] and *riskantní* [risky] that are normally considered abstract, have high association strength along these dimensions.

We have demonstrated that the individual scales used to capture the socio-semantic dimensions are reliable, with strong agreement across raters, which is important for our new dimensions of LOCATION and POLITICAL. Additionally, our analyses revealed that for our word stimuli, there is a clear relationship to existing norms for translation equivalent items in English for the dimensions of GENDER and VALENCE, providing more support for their validity, as was the case in Scott et al (2019) and Warriner et al (2013). Our analyses of how the dimensions relate to each other has provided new empirical insights, revealing a number of interesting relationships that highlight how some concepts may co-vary in terms of their socio-semantic representations, i.e. concepts that are more liberal may also have associations to femininity, urban environments, positive emotions and younger age groups, whereas more conservative concepts may be also have associations with masculinity, rural environments, negative emotions and older age groups.

By providing data not only for words, but also colour and grayscale images, we have been able to assess the extent to which different versions representing the same underlying concept may vary in the way they represent socio-semantic meaning. Our analyses suggest that there is generally strong alignment across the different stimuli versions, but we highlight that not all items may be represented similarly, indicating that colour may bias the representation, or even lead to processing the item as something conceptually different. Thus, we emphasise the importance of considering how concepts may not align across different presentation modalities or versions, if only relying on norms derived from word labels.

Nevertheless, there are still a number of important directions for future research that will help improve the utility of socio-semantic norms. We acknowledge that our participant sample, although relatively large, is heavily skewed and restricted to a subset of the more general population. Specifically, we collected the data for our norms from a predominantly female sample of university students of humanities, aged between 18-30, with more liberal attitudes and urban affiliation. Although other norming studies have also used a similar demographic sample (e.g. sensory modality norms by Speed & Brysbaert, 2022), the collection of norms where the dimensions relate to socially meaningful dimensions, such as gender, location, political and age, may be influenced by the individual participants socio-demographic profile. The extent to which certain demographic groups (e.g. female vs male or young vs old) may differ remains an open research question. Although the primary focus of the current paper is to establish and introduce the first large-scale socio-semantic norms, we are currently collecting more data from more diverse samples of participants to address the question of whether socio-semantic representations are stable or variable across different socio-demographic groups.

Open Practices Statement

The anonymised raw data and summary data from this paper, as well as the code used to run the data processing, visualisation and analyses can be found on the Open Science Framework at:
https://osf.io/pv9md/?view_only=ccbd13b7afb141628b51e0d0a83f9037

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.

References

Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. *Journal of Cognition,* 3(1).

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4), 130-174.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, 29, 120-153.

Brand, J., Hay, J., Clark, L., Watson, K., & Sóskuthy, M. (2021). Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics*, 88, 101096.

Brand, J., Preininger, M., Kříž, A. & Ceháková, M. (under review). Feminine fox, not so feminine box: Constraints on linguistic relativity effects for grammatical and conceptual gender. *Language and Cognition*.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904-911.

Christensen R (2023). Ordinal: Regression Models for Ordinal Data. R package version 2023.12-4,

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Diveica, V., Muraki, E. J., Binney, R. J., & Pexman, P. M. (2024). Socialness effects in lexical–semantic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 50*(8), 1329–1343.

Diveica, V., Pexman, P. M., & Binney, R. J. (2023). Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods,* 55(2), 461-473.

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.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, 50, 1116-1124.

Ferré, P., Guasch, M., Stadthagen-Gonzalez, H., & Comesaña, M. (2022). Love me in L1, but hate me in L2: How native speakers and bilinguals rate the affectivity of words when feeling or thinking about them. *Bilingualism: Language and Cognition*, 25(5), 786-800.

Grühn, D., & Smith, J. (2008). Characteristics for 200 words rated by young and older adults: Age-dependent evaluations of German adjectives (AGE). *Behavior Research Methods*, 40, 1088-1097.

Haspelmath, M. & Tadmor, U. (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Berlin, New York: De Gruyter Mouton.

Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PloS one*, 14(2), e0210793.

Imbault, C., Titone, D., Warriner, A. B., & Kuperman, V. (2021). How are words felt in a second language: Norms for 2,628 English words for valence and arousal by L2 speakers. *Bilingualism: Language and Cognition*, *24*(2), 281-292.

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., & Šindlerová, J. (2021). SYN2020: a new corpus of Czech with an innovated annotation. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021*, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24 (pp. 48-59). Springer International Publishing.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, 37(2), 214-233.

Labov, W. (2011). *Principles of linguistic change, volume 3: Cognitive and cultural factors (Vol. 3)*. John Wiley & Sons.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour,* 4(10), 1021-1028.

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.

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., ... & Ünal-Logacev, Ö. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words?. *Behavior Research Methods*, 48, 1154-1177.

Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolík, F., Butcher, M., Chondrogianni, V., ... & Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *PloS one*, 14(8), e0220611.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52, 1271-1291.

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., ... & Sczesny, S. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*, 46, 841-871.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A. L., ... & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45, 169-177.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning (No. 47)*. University of Illinois press.

Pexman, P. M., Diveica, V., & Binney, R. J. (2023). Social semantics: the organization and grounding of abstract concepts. *Philosophical Transactions of the Royal Society B*, 378(1870), 20210363.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. *Behavior Research Methods*, 51, 453-466.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

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

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.

Sehyr, Z. S., Caselli, N., Cohen-Goldberg, A. M., & Emmorey, K. (2021). The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2), 263-277.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.

Speed, L. J., & Brysbaert, M. (2024). Ratings of valence, arousal, happiness, anger, fear, sadness, disgust, and surprise for 24,000 Dutch words. *Behavior Research Methods*, 56(5), 5023-5039.

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.

Trudgill, P. (1974). *The social differentiation of English in Norwich (Vol. 13)*. CUP archive.

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.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Veríssimo, J. (2021). Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models. *Bilingualism: Language and Cognition*, 24(5), 842-848.

Walker, A., & Hay, J. (2011). Congruence between ‘word age’ and ‘voice age’ facilitates lexical access. *Laboratory Phonology*, 2(1).

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1), 106.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45, 1191-1207.

Wilson Black, J., Brand, J., Hay, J., & Clark, L. (2023). Using principal component analysis to explore co-variation of vowels. *Language and Linguistics Compass*, 17(1), e12479.

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2024). Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, 56(3), 1640-1655.

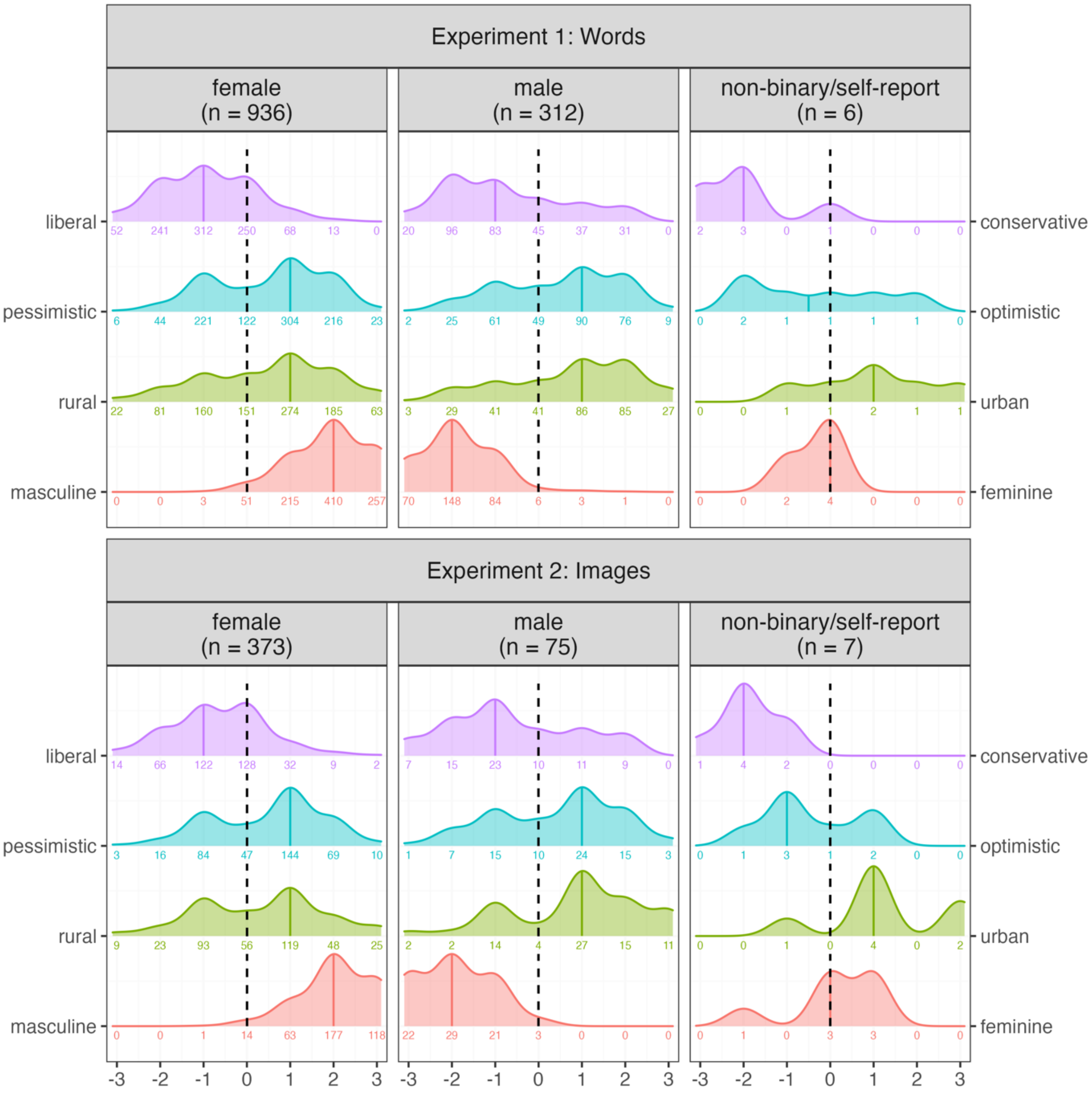
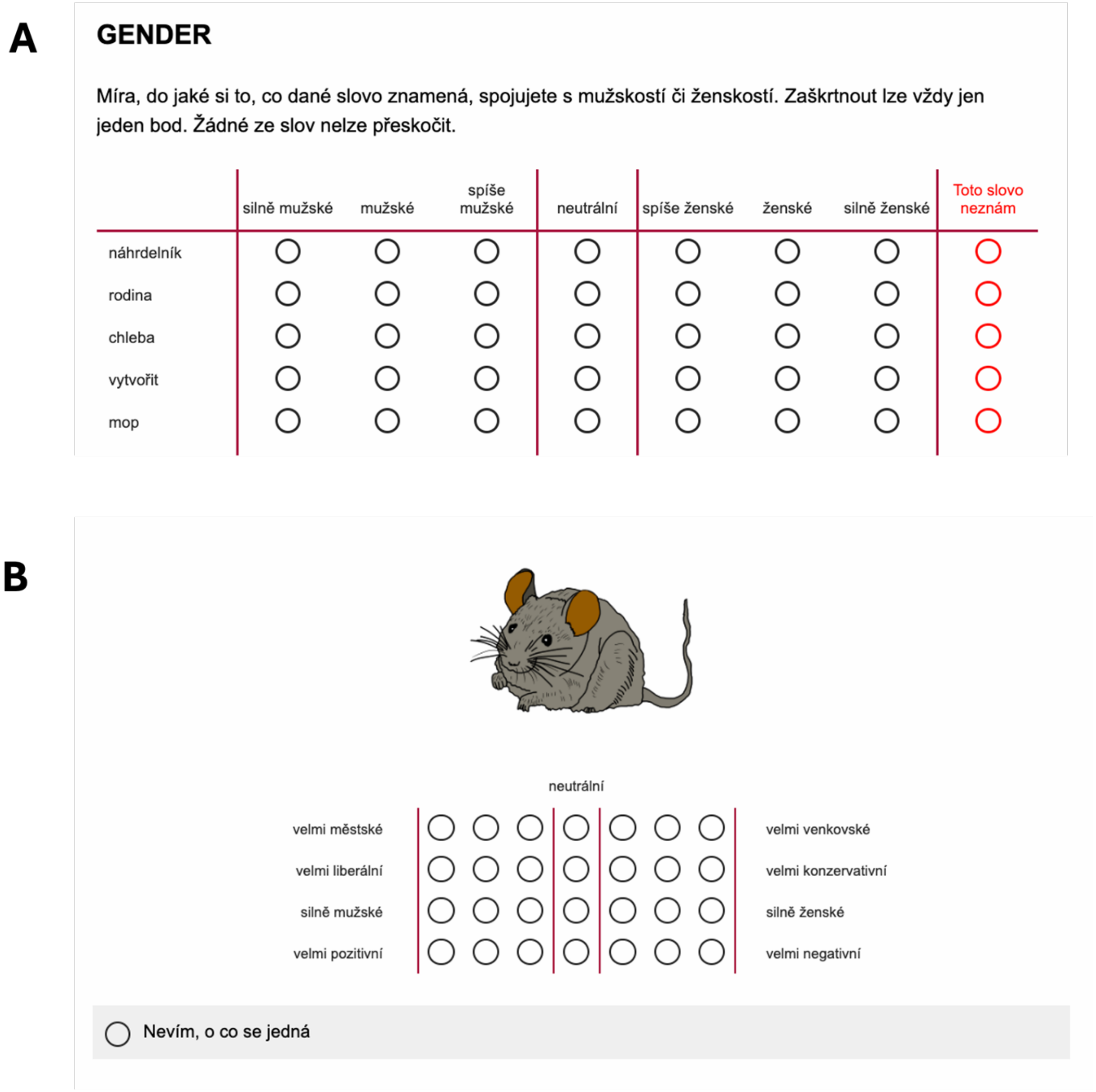


Figure 1: Demographic profile of participants from Experiments 1 and 2. Facets are used to separate the data based on categorical gender, with the number of participants given in brackets. The y-axis separates the different demographic questions. The values on the x-axis represent a numeric transformation of the 7-point Likert scale, i.e. -3 represents very liberal/pessimistic/rural/masculine, 3 represents very conservative/optimistic/urban/feminine and 0 represents the neutral mid-point. Counts of participants who selected each scale point are given as numbers under each density plot, e.g. In Experiment 1, there were 52 out of the 936 females who identified as very liberal.



GENDER

Míra, do jaké si to, co dané slovo znamená, spojujete s mužskostí či ženskostí. Zaškrtnout lze vždy jen jeden bod. Žádné ze slov nelze přeskočit.

	silně mužské	mužské	spíše mužské	neutrální	spíše ženské	ženské	silně ženské	Toto slovo neznám
náhrdelník	○	○	○	○	○	○	○	○
rodina	○	○	○	○	○	○	○	○
chleba	○	○	○	○	○	○	○	○
vytvořit	○	○	○	○	○	○	○	○
mop	○	○	○	○	○	○	○	○

				neutrální				
velmi městské	○	○	○	○	○	○	○	velmi venkovské
velmi liberální	○	○	○	○	○	○	○	velmi konzervativní
silně mužské	○	○	○	○	○	○	○	silně ženské
velmi pozitivní	○	○	○	○	○	○	○	velmi negativní

○ Nevím, o co se jedná

Figure 2: Examples of the rating procedure used in Experiment 1 (image A) and Experiment 2 (image B).

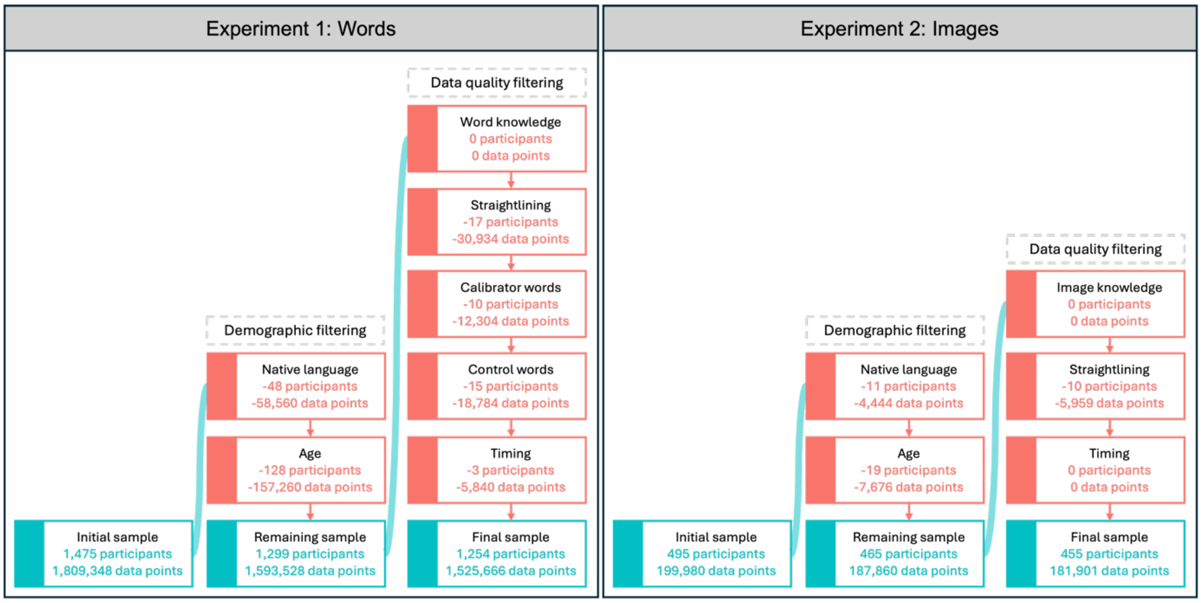


Figure 3: Filtering procedure applied to the data from Experiment 1 and 2. Turquoise boxes contain summaries of the data that remains after each filtering block, red boxes contain summaries of the data removed during each filtering step.

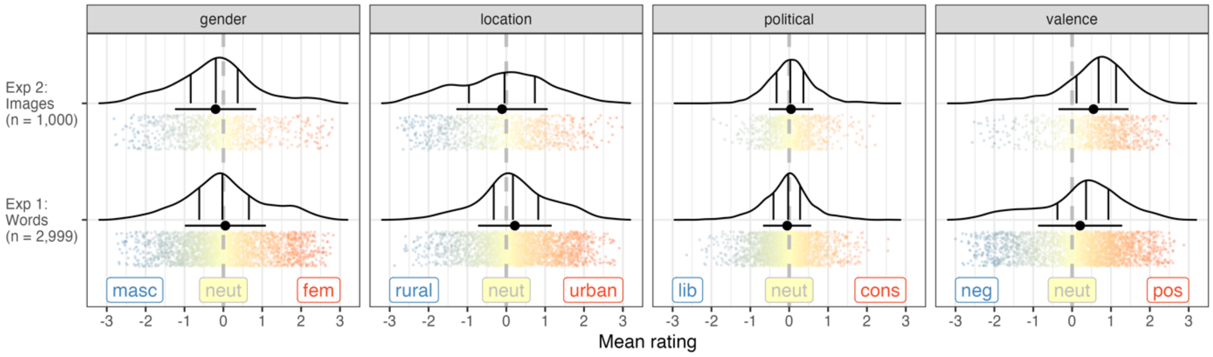


Figure 4: Distribution of mean ratings for the GENDER*,* LOCATION*,* POLITICAL *and* VALENCE *dimensions for Experiment 1 and 2. Kernel density estimates are shown with 25%, 50% and 75% quantiles marked by solid vertical lines, with the dashed line representing a value of 0 (neutral). The black point with a horizontal line represents the mean and standard deviation of the ratings. Each word is represented by a point, with more red/blue colours indicating stronger association towards a specific side of the scale and yellow points associating more with neutral ratings.*

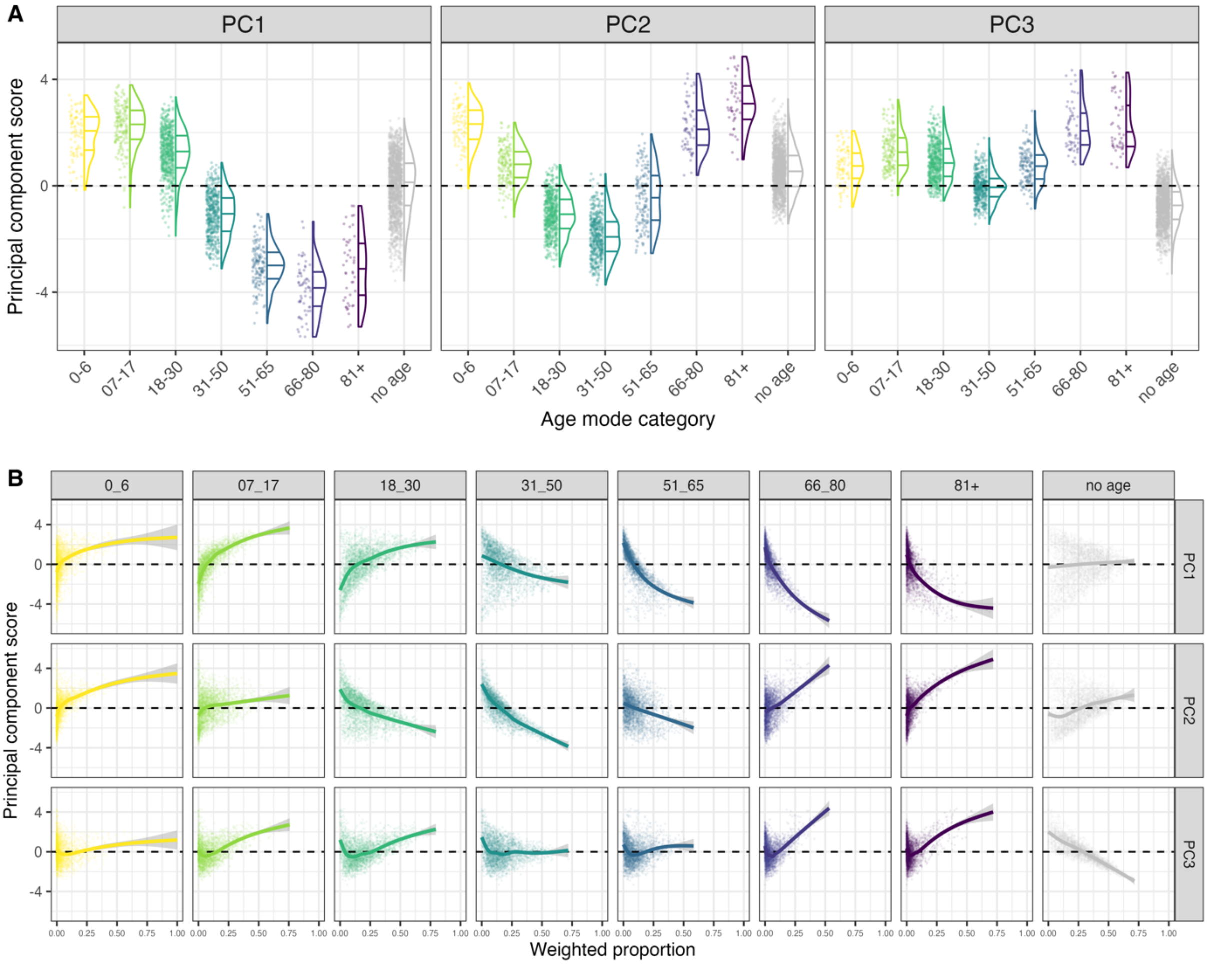


Figure 5: Visualisation of the results from the PCA for the age dimensions. A: Distribution of PC scores (y axis) based on the mode age of each word (x axis) for each of the three PCs (facets). B: The weighted proportions for each word (x axis) based on the age categories (top facets) and their relationship to PC scores (y axis) with a gam smooth by PC (side facets).

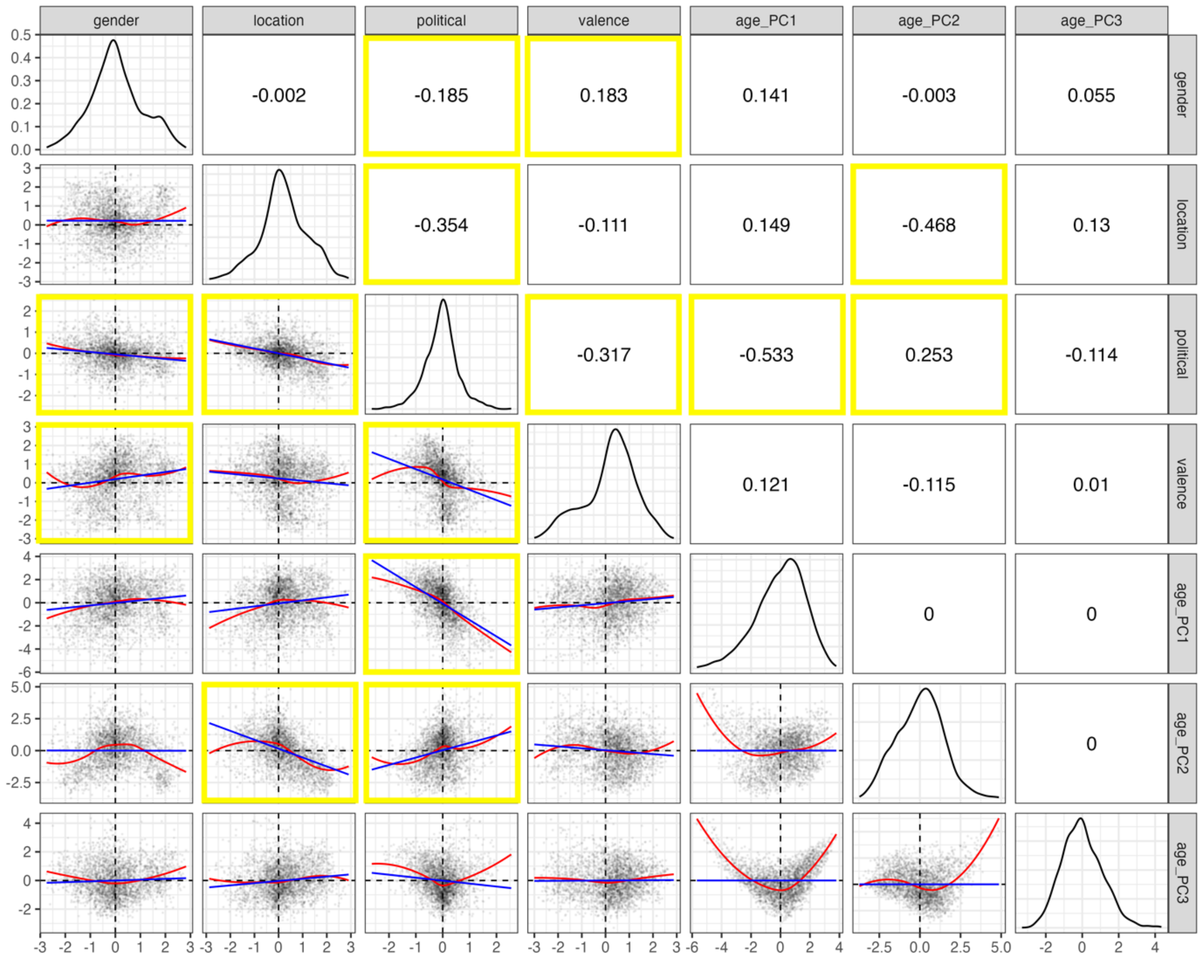


Figure 6: Correlation matrix of all the dimensions in the word ratings dataset. The lower left plots are scatter plots of the data fitted with a linear (blue) and loess (red) fit, with the data from the dimension in the top facets on the x-axis and the data from the right facet on y-axis. The diagonal plots show kernel density distributions of the variables. The upper right plots give the Pearson correlation coefficients between two dimensions. Notable correlations (r >|.18|) are highlighted by yellow outlines.

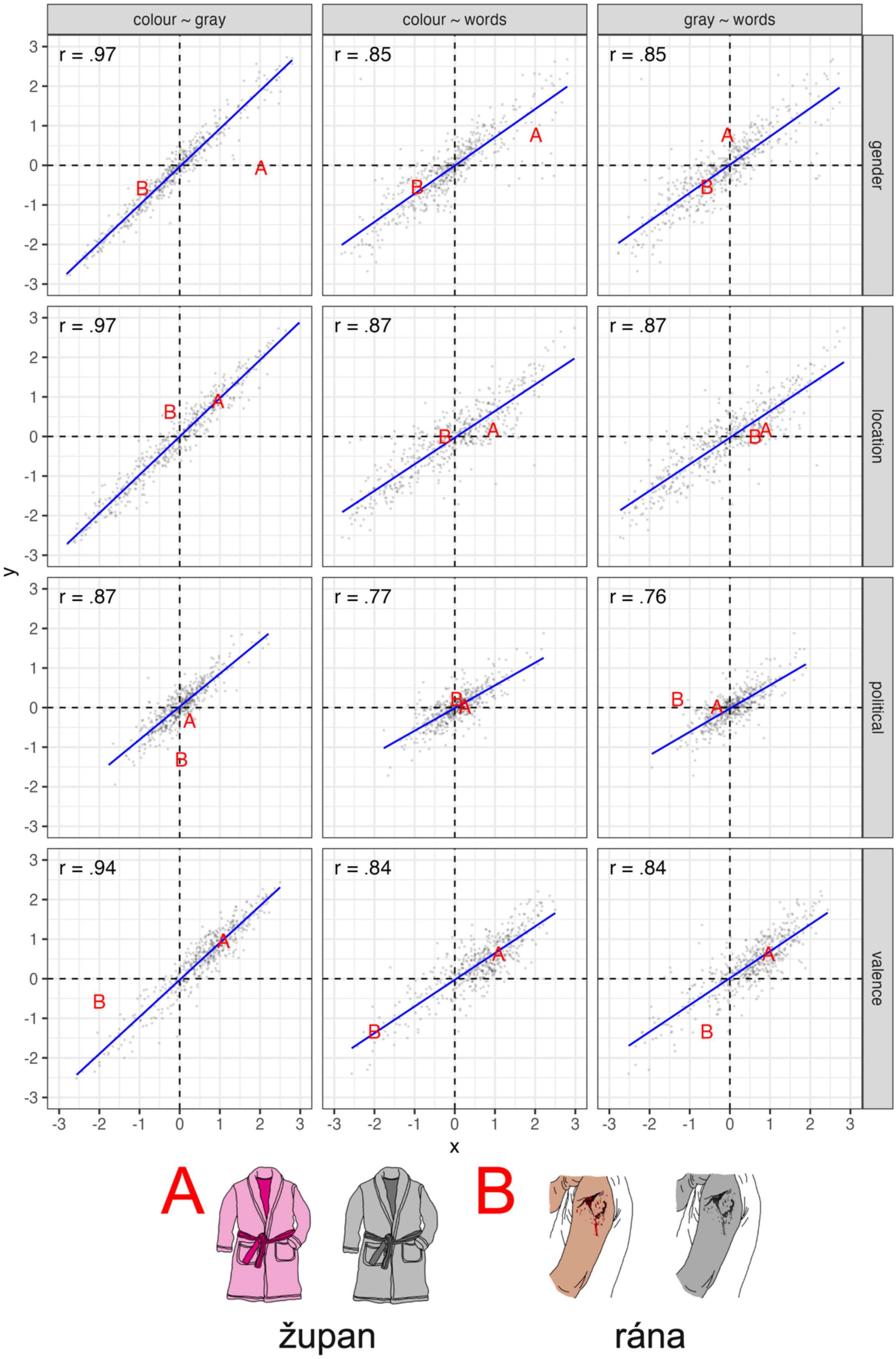


Figure 7: Comparisons between the colour, grayscale and word versions of the items from the Multilingual Picture Database (Duñabeitia et al., 2022). The facets on the top divide the data by stimuli comparison, with the mean ratings values given on the x and y axis respectively, e.g. colour~gray has ratings for items in colour on the x axis and grayscale items on the y axis. The facets on the right divide the data by the dimensions. The Pearson's correlation coefficients are given in the top left corner of each facet, with a linear fit to the data added in blue. Two example stimuli are represented on the plots as red letters - A (dressing gown) and B (cut/wound) - with the colour, grayscale and Czech word versions shown at the bottom of the plot.

Table 1: Summary of the number of responses to items across the two experiments.

Experiment	Median N	Range N	Median N known	Range N known	Mean prop known	Range prop known
Exp 1: Words	37	[28,96]	37	[11,96]	.997	[.35,1]
Exp 2: Images	45	[35,52]	45	[10,52]	.978	[.26,1]

Table 2: Items with the most extreme values for each of the dimensions available in the dataset. Items are given in Czech with their associated values (means or PC scores) in brackets. See the supplementary materials for English translation equivalents.

GENDER	Masculine	*otec (-2.74), varlata (-2.71), táta (-2.69), muž (-2.67), Vladimír (-2.67), penis (-2.62), mužský (-2.61), strýc (-2.59), Rudolf (-2.55), Adam (-2.53)*
	Feminine	*Božena (2.83), princezna (2.75), vagina (2.75), Sofie (2.7), Adéla (2.68), podprsenka (2.68), matka (2.65), Karolína (2.64), Viktorie (2.63), žena (2.62)*
LOCATION	Rural	*venkov (-2.88), orat (-2.62), venkovská (-2.59), vesnice (-2.59), farmářka (-2.59), traktor (-2.58), farma (-2.55), chalupa (-2.4), stodola (-2.39), venkovský (-2.38)*
	Urban	*Praha (2.91), velkoměsto (2.78), metropole (2.78), mrakodrap (2.77), město (2.74), Berlín (2.74), New York (2.72), fastfood (2.71), Vídeň (2.68), Londýn (2.68)*
POLITICAL	Liberal	*liberalismus (-2.64), liberální (-2.3), transgender (-2.24), bisexuálka (-2.18), multikulturalismus (-2.15), bisexuál (-2.14), lesba (-2.13), transsexuál (-2.12), veganka (-2.12), transsexuálka (-2.07)*
	Conservative	*konzervativní (2.55), konzervatismus (2.54), monarchie (2.04), homofobie (1.97), církev (1.92), křešťanství (1.89), bible (1.88), království (1.85), šlechta (1.84), staromódní (1.83)*
VALENCE	Negative	*znásilnění (-3), znásilnit (-2.84), rakovina (-2.82), terorismus (-2.82), násilník (-2.78), terorista (-2.73), pedofil (-2.68), vrah (-2.67), zmrd (-2.67), nacismus (-2.65)*
	Positive	*šťastný (2.88), zdravá (2.63), šťastná (2.62), láska (2.59), smát se (2.59), spokojenost (2.59), milovat (2.54), radost (2.53), svoboda (2.53), štěstí (2.51)*
AGE PC1	Old	*penze (-5.68), stáří (-5.66), seniorka (-5.49), děda (-5.4), babička (-5.4), senior (-5.31), stará (-5.3), starý (-5.23), alzheimer (-5.19), stárnutí (-5.17)*
	Young	*beďar (3.79), gymnázium (3.73), panictví (3.69), dospívání (3.6), selfie (3.52), opisovat (3.51), akné (3.48), škola (3.47), výtvarka (3.46), drzý (3.42)*
AGE PC2	Middle age	*statistik (-3.72), rozvod (-3.58), účetní (-3.53), podnikání (-3.47), dozorce (-3.37), architektka (-3.37), popelářka (-3.35), byznys (-3.34), exekutorka (-3.33), manažer (-3.3)*
	Young/old	*hřbitov (4.86), rakev (4.8), kremace (4.8), starý (4.51), zemřít (4.47), stará (4.4), pohřeb (4.4), smrt (4.34), senior (4.21), demence (4.12)*
AGE PC3	No age	*přehrada (-3.31), Litva (-2.89), tsunami (-2.78), špičatý (-2.74), Lotyšsko (-2.61), Irsko (-2.61), široká (-2.57), voda (-2.52), hmyz (-2.51), hovězí (-2.5)*
	Old	*senior (4.34), rakev (4.26), stáří (4.16), seniorka (4.14), starý (4.13), stará (4.11), hřbitov (4.08), penze (3.94), děda (3.83), kremace (3.72)*