Quantifying the Socio-semantic Representations of Words

Mikuláš Preininger (mikulas.preininger@ff.cuni.cz)
Faculty of Arts, Charles University, Czech Republic

James Brand (james.brand.ac@gmail.com)
Faculty of Arts, Charles University, Czech Republic

Adam Kříž (adam.kriz@ff.cuni.cz)
Faculty of Arts, Charles University, Czech Republic

Abstract

Quantifying the meaning of a word is a complex challenge. Humans can encode semantic information along a large and diverse range of semantic dimensions for any given word. Whilst a number of studies have applied a range of techniques to quantify word meaning along specific dimensions, little work has focussed on the socio-semantic dimensions of meaning. Here, we present data that quantifies the socio-semantic representations of 2,700 Czech words along the dimensions of gender, location, political, valence and age. We also demonstrate the utility of the data set by calculating an estimate of socio-semantic similarity between all words, which can be used to identify words that are either proximally close or distant in socio-semantic space.

Keywords: semantics; concepts; norms; similarity; Czech

Introduction

The ability to quantify the meaning of words has been a long-standing goal for the cognitive sciences. Since the early work by Osgood, Suci and Tannenbaum (1957), there has since been a wide range of different approaches used by researchers, which have been focused on obtaining measurements of word meaning. These approaches vary from high-dimensional semantic spaces derived from text corpora (e.g. word embeddings, Mikolov et al., 2013), to unidimensional normative ratings from human participants, which focuses on a theoretically motivated aspect of meaning (e.g. concreteness, Brysbaert et al., 2014).

Whilst the word embedding approach has become exceptionally popular in recent years, given it can be applied to many different languages and different types of linguistic data (e.g. Grave et al., 2018; van Paridon & Thompson, 2021), there still remains clear benefits for normative rating approaches. For example, the researcher can clearly define a specific dimension of word meaning to be quantitatively normed for a list of words, measuring participant associations between the words and the properties of the dimension. The resulting data set can then be used to test theoretical predictions specifically related to the dimension (e.g. abstract words are processed slower than concrete words, Brysbaert et al., 2014). Thus, the approach is particularly appealing when addressing questions related to definable dimensions of meaning.

The variety and size of normative data sets now available in a number of different languages, highlights the importance of the approach for psychologists, linguists and cognitive scientists more broadly. These norms capture meaning of a theoretically defined construct, either unidimensionally (e.g. iconicity, Winter et al., 2017) or multidimensionally (e.g. sensorimotor strength across 11 different dimensions, Lynott et al., 2020). However, there has only been a limited amount of attention given to quantifying social dimensions of meaning, or in other words socio-semantic representations. For instance, whether people associate the meaning of a word towards a particular gender, location, political ideology or age group.

This is surprising given the extensive literature from sociolinguistics that has demonstrated the important role of socially encoded information in language production and perception (for a recent review, see Hay, 2018). Thus, if words encode socio-semantic information, then the norming approach used for dimensions such as concreteness, should also be a valuable tool to quantify socio-semantic dimensions. For instance, Scott et al. (2019) presented the first large-scale study investigating how words are rated in terms of their association to gender (i.e. feminine – neutral – masculine), with a similar approach reported in Lewis et al (2022). Ratings of whether a word is related to young or old age have also been studied, albeit on a much smaller scale (e.g. Grühn & Smith, 2008). This norming approach differs from corpus derived estimates of gender (e.g. Sap et al., 2014), where the quantification is based on the frequency of usage by specific socio-demographic sub-group, i.e. which words are likely to exhibit similar or dissimilar production frequencies when comparing texts written by males or females. However, the focus of the norming approach is to quantify the meaning representation, which is distinct from frequency of usage, therefore offering a unique insight into how people associate a word to specific dimensions of meaning (e.g. the word *boyfriend* might be used more by females, but the meaning representation is more likely to be associated to males).

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

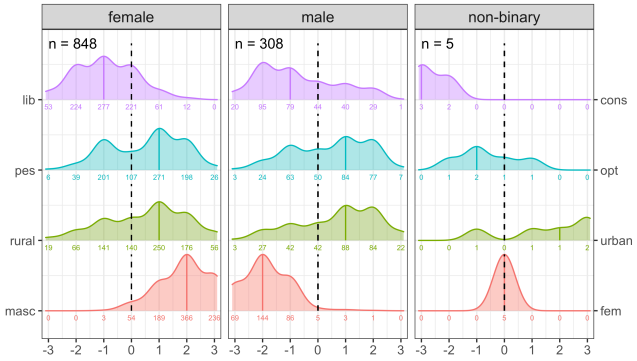

Figure 1: Demographic profiles of participants. Values of -3 on the x-axis correspond to liberal/pessimistic/ rural/masculine, whereas values of 3 correspond to conservative/optimistic/urban/feminine. The neutral midpoint is represented by the dashed line.

The aim of the present paper is to provide the first large-scale quantification of socio-semantics, providing normative data for 2,700 Czech words across 5 different socially meaningful dimensions (Study 1). We also demonstrate the practical utility of the norms by calculating a measurement of socio-semantic similarity, so that clusters of similar (or dissimilar) words can be identified (Study 2).

Study 1: Socio-semantic norms

The primary aim of this study was to establish a large data set of subjective ratings from a population of Czech speaking adults, capturing 5 distinct dimensions of socio-semantic meaning (GENDER, LOCATION, POLITICAL, VALENCE and AGE). We present several aggregated variables that provide individual word levels norms across the dimensions. Finally, we explore the correlations that exist between the dimensions to understand how words might pattern together in terms of their associated representations.

Method

Participants In total 1,161 participants took part in the study (848 identified as female, 308 as male and 5 as non-binary), who were recruited from a university wide student database at Charles University, in addition to recruitment via Prolific (https://www.prolific.co/). All participants were aged between 18-30 years old (M = 21.8, SD = 2.3), were native (or highly proficient) speakers of Czech. Ethics approval was granted by the ethics commission of the Faculty of Arts, Charles University.

Stimuli Our study comprises a list of 2,700 Czech words. Items were chosen to represent a broad range of diverse semantic domains (such as occupations, religion, tools, personal traits etc.), spanning different parts-of-speech (1,603 nouns, 766 adjectives, 331 verbs) and lexical

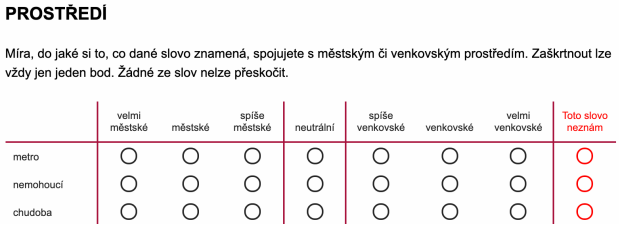

Figure 2: Screenshot of the LOCATION rating scale. The scale ranges from (left to right): *very urban, urban, slightly urban, neutral, slightly rural, rural, very rural.* The final red column is used when the word meaning is not known.

frequencies (derived from the Syn-v9 corpus of written Czech, Křen et al., 2021). As Czech has grammatical gender (GG), we decided to include both masculine and feminine variants of nouns and adjectives whenever possible (e.g. the noun [DIRECTOR] can be *ředitelka (fem.)* or *ředitel (masc.)*; the adjective [STRONG] can be *silná (fem.)* or *silný (masc.)*.[1] The fully annotated word list is available in the online supplementary materials.

The list of 2,700 words was pseudo-randomly divided into 27 separate 100-word subsets. Each subset contained approximately the same number of words from each part-of-speech category, we also controlled the distribution of GG words by ensuring no subset contained both GG variants of the same word and the number of feminine/masculine adjectives and nouns was roughly comparable across lists. All subsets were further complemented with four phonotactically plausible non-words (e.g. *tontota*) and a calibrator word for each socio-semantic dimension (the first word participants rated, chosen on the basis of a pilot experiment, e.g. *metro* [SUBWAY] was chosen for LOCATION as it was reliably rated as very urban. These words were used as a quality check on participant data.

Procedure The data were collected online via a questionnaire designed using Qualtrics. The questionnaire consisted of brief instructions followed by a short socio-demographic questionnaire which involved self-assessment of the participant's age, gender identity, education, and native language. We also used 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 these questions contained a neutral midpoint. See Figure 1 for visualisation of these responses.

Participants were then asked to rate each of the words from one of the 27 subsets (i.e. 100 words and 5 control words), specifically by how they associated the word according to each of the following dimensions: GENDER *(very masculine - very feminine)*; LOCATION *(very urban - very rural)*;

[1] Although it is outside the main aims of the present paper to analyse in detail how part-of-speech, GG and participant demographic differences affect the normative ratings, this is a topic of ongoing research, see discussion section for more details.

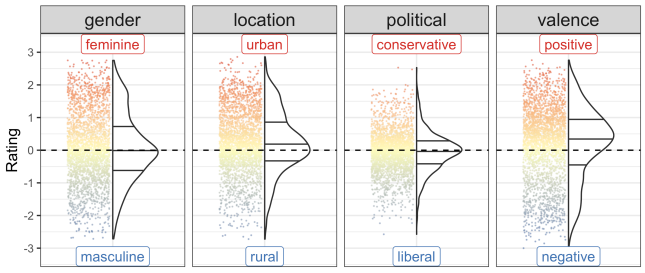

Figure 3: Distribution of mean ratings for the GENDER, LOCATION, POLITICAL, VALENCE dimensions. Kernel density estimates are shown with 25%, 50% and 75% quantiles marked by solid horizontal lines, with the dashed line representing a value of 0 (neutral). Each word is represented by a point, with more red/blue colours indicating stronger association towards a specific side of the scale.

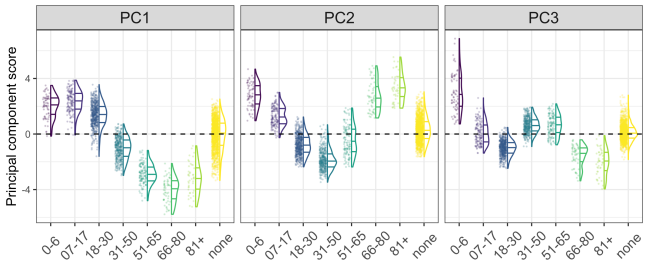

Figure 4: Distribution of AGE PC scores (y-axis) for the 3 principal components (PC1: young/old; PC2: middle aged, PC3: childhood/old age). Each point represents a word categorised by the age category with the highest proportion of ratings (x-axis).

POLITICAL *(very liberal - very conservative)*; VALENCE *(very positive - very negative)*; and AGE (divided into categories of *0-6, 7-17, 18-30, 31-50, 51-65, 66-80, and 81+ years*). Participants were presented with one dimension at a time, which contained all the words from one of the subsets. Nouns and adjectives were shown in nominative singular, whereas verbs in infinitive form. The order of presentation for dimensions and words was randomised for each participant (apart from the calibrator word, which was always presented first). All dimensions (apart from AGE) were rated using 7-point Likert scales, each with a neutral midpoint and the option to skip a word if the meaning was not known. See Figure 2 or the supplementary materials for a working example of the experiment (available only in Czech).

For the AGE dimension, participants could choose one, multiple, or none of the options. This differs from the design used for the other dimensions because we wanted to assess how words can be related to different age categories found across the human lifespan. This approach provides a more nuanced measurement of age association in comparison to using more linear scale (e.g. *young – old* as used in Grühn & Smith, 2008), which would not be able to distinguish between distinct age categories as clearly. This also allows us to assess whether a word is associated with a single (e.g. *kindergarten* = 0-6), several (e.g. *basketball* = 7-17 and 18-30) or none of the age categories (e.g. *bamboo*).

Data processing

The median number of participants who were assigned to each subset of 100 words was 42 (range = 33-57) and only 6 words out of the total 2,700 were rated as unknown by > 20% of the participants. Based on the ratings provided by each of the participants we calculated several aggregated statistics which provide each individual word with a single interpretable value relating to each of the 5 dimensions. The dataset is available in the supplementary materials.

Proportions To maintain the multidimensionality of the rating scales, we first calculated proportions of responses to each point on the scale. This was done by recoding the participant responses as either 1 (for the selected point) or 0 (for all other points). From these values we can thus calculate a proportion value for each word by summing the response values and dividing by the number of participants who rated the word. This produces values ranging between 0 (no ratings) and 1 (all participants chose the same point on the scale). This meant that for the dimensions of GENDER, LOCATION, POLITICAL, VALENCE each point along the 7-point Likert scale would have a proportion variable for each of the words (i.e. there would be 7 distinct variables, each with a proportion value calculated).

For AGE the calculation of proportions was modified to take into account the fact that multiple options could be selected by the participant for an individual word. This meant that a weighted rating value was calculated for each participant, based on the number of age categories selected for a word (1/n_selected_categories), i.e. if a participant selected 2 age categories for a word, each category would have a weighted rating of 0.5, if 1 category was selected then the value would be 1, if all 7 categories were selected the value would be 1/7. From the weighted values we could then calculate a weighted proportion for each word, for each of the age categories.

Descriptive statistics For the dimensions of GENDER, LOCATION, POLITICAL, 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. From this data we calculated the mean and SD for each of the words. See Figure 3 for visualisation of the data. For AGE we used the proportions data to obtain the age category for each word with the highest proportion of ratings across the possible age options (which could only be a single option out of *0-6, 7-17, 18-30, 31-50, 51-65, 66-80, 81+* or *none*), providing a categorical measure of age association.

Latent means In order to preserve the ordinal nature of the Likert scales used for the GENDER, LOCATION, POLITICAL, VALENCE dimensions, we followed the guidance from Taylor et al (2021) and modelled the participant responses using Cumulative Link Mixed-effects Models. These models

account for variation that is introduced from the participant response biases and thus provide a more accurate estimation of a normative value for each word. We modelled 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) simply by random intercepts for word and participant.[2] From this we were able to extract the random intercepts for the effect of word, providing us with a numeric estimate of the latent mean for each word.

PCA Age As the only measurement of age association we have so far is categorical, we also decided to compute linear estimates for AGE. To do this, we used the proportion data, whereby each age category is a distinct variable with numeric values for each word and ran a Principal Component Analysis (PCA) on that multidimensional space. This approach not only allows us to reduce the space down to composite variables (principal components, PCs), but within these variables we can also obtain a score for each word, which can be used as an estimate for age association in relation to the variables loaded on each of the PCs (see Brand et al., 2021 for details).

The PCA resulted in 3 main PCs (see Figure 4). PC1 (accounting for 38.7% of the variance) distinguishes between words associated with young age groups (*0-6, 7-17 and 18-30)* and old age groups (*51-65, 66-80 and 81+*), with associations to middle age (*31-50*) and no age associations being in between. PC2 (accounting for 27.5% of the variance) distinguishes between words associated with middle age (*31-50*) and words at the youngest and oldest ages (*0-7, 7-17 and 66-80, 81+*). PC3 (accounting for 14.5% of the variance) distinguishes between words associated with childhood (*0-7*) and old age (*66-80 and 81+*).

Analysis

We first wanted to establish whether there was a substantial difference between the ratings calculated by the mean and the latent mean for the GENDER, LOCATION, POLITICAL, VALENCE dimensions. This was assessed by a pairwise correlation, which showed that the ratings were almost perfectly correlated (all Pearson's r values > 0.98). Given the benefits of using the latent mean put forward by Taylor et al. (2021), we decided to use the latent mean values instead of the mean for all subsequent analyses.

In order to explore the relationships that exist between the variables, we ran a series of exploratory correlation analyses. As these were exploratory and had a large number of multiple comparisons, we do not place strong emphasis on significance testing, but instead focus on describing the main trends found from the analysis. See Figure 5 for full results.

The strongest correlation (r = -.562) came between POLITICAL and PC1 for age, indicating that words rated as more conservative were also more likely to be associated with older age (and more liberal words with younger ages).

[2] R syntax for models using the `ordinal` library: `clmm(rating~ 1 + (1 | word) + (1 | participant), data = dimension_data, link = "probit")`

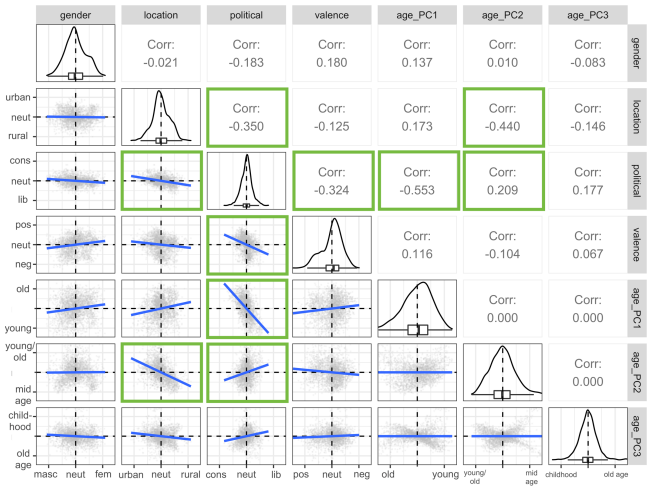

Figure 5: Relationships between the socio-semantic dimensions (using latent mean values) and the 3 PCA derived age variables. The lower portion shows a linear regression fit to the data. The upper portion reports the Pearson's correlation co-efficient. The plots along the diagonal give the kernel density estimates for the distribution of ratings. Green boxes highlight correlations >.2.

POLITICAL was also correlated with PC2 for age (r = .222, indicating that more conservative words are also associated with middle age ratings), LOCATION (r = -.353, indicating that more conservative words are associated with rural ratings) and VALENCE (r = .288, indicating that more conservative words are associated with negative ratings). LOCATION was also correlated with PC2 for age (r = -.443), indicating that more urban words are associated with middle age.

Study 2: Socio-semantic similarity

The aim of the second study was to investigate whether the data from the previous study can be used to capture socio-semantic similarities (or dissimilarities) between words. This first required a representation of the multidimensional socio-semantic space to be generated. From this representational space, we then aimed to estimate a measure of similarity between all words (using similar methods as used by Wingfield & Connell, 2021), to see if the rating data can proximately cluster words together meaningfully.

Method

Data In order to most accurately preserve the multidimensionality of the available data, we selected the variables from the previous study that represented the proportions of participants who selected each value on each of the dimension's rating scales. This meant that for each of the 5 socio-semantic dimensions, we had 7 different variables

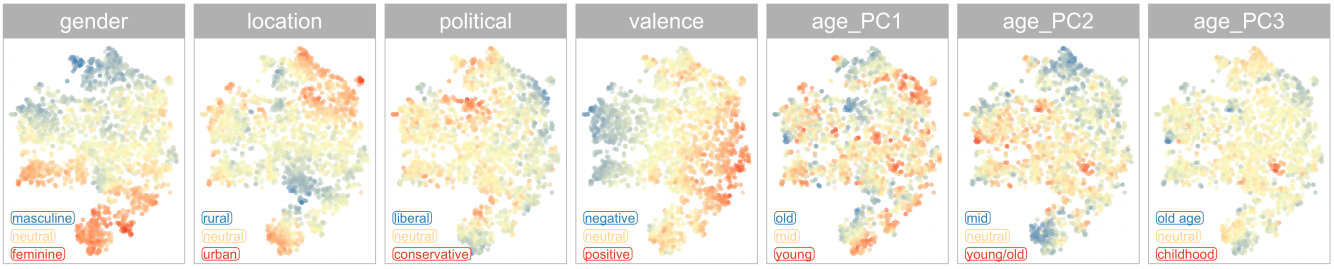

Figure 6: Visualisation of the t-SNE mapping. Each facet uses colour to show the latent mean rating or PC score of each individual point for the different socio-semantic dimensions (representing an individual word in the norming data set). The more blue/red the points are, the stronger the association is towards the respective sides of the scale, e.g. very red points would have a very feminine rating in the first facet.

representing the participant aggregated proportions for each word (e.g. for LOCATION the 7 dimensions were *very urban, urban, slightly urban, neutral, slightly rural, rural, very rural*). This resulted in a 35 dimension × 2,700 word data set. We chose this approach over using the unidimensional variables (e.g. means, as in Wingfield & Connell, 2021) as a richer multidimensional space will preserve the differences in ratings from each of the points on the scales, eliminating concerns about word ratings with large variance across participants, i.e. unidimensional variables do not encode information about variation (such as SD), whereas the proportion variables do.

Analysis

Visualisation The first step in our analysis was purely for exploratory purposes. In order to visualise such a high dimensional data set, we created a t-SNE mapping (Maaten & Hinton, 2008), using the `Rtsne` package, (Krijthe, 2015). As all of our variables were based on proportions (i.e. all values ranged from 0 to 1), we did not normalise or run a PCA on the data before running the t-SNE. We used a perplexity parameter of 30 and a theta value of 0.

In order to understand how the t-SNE mapping distributes words in the 2-dimensional space, we present in Figure 6, the same t-SNE mappings, but facetted by the different socio-semantic dimensions. Each facet colours the individual points based on their latent mean values for GENDER, LOCATION, POLITICAL, VALENCE and PC scores for the 3 different AGE PCs. From this we can see that there is a clear distinction in the space based on the GENDER (vertical) and VALENCE (horizontal) dimensions, with the other dimensions distributed within the space (e.g. urban, liberal and middle aged related words are all located at the extremes of the vertical). An interactive version of the t-SNE is available in the supplementary materials, where individual word neighbours can be viewed.

Although the visualisation offers an alternative to the correlation analysis for understanding how words might cluster together based on their ratings, we also need to establish a quantitative measure of similarity and dissimilarity between the words, without reducing the multidimensional space.

Cosine similarity Following the approach used by Wingfield and Connell (2021), we created a cosine similarity matrix between all possible combinations of words in our data set (resulting in over 7 million pairwise similarities). The cosine similarity was calculated using the 35-dimensional vector of socio-semantic proportions. Similarities between any two words in the multidimensional space can have a value between 0 (maximally different) and 1 (maximally similar). The complete set of pairwise distances between the words can be found in the supplementary materials.

In order to provide a descriptive overview of the calculated distances, we will provide a few representative examples from the data set. Examples of word pairs with the highest similarity (cosine distance > .99) were largely to do with non-sexual body parts and were largely neutral across all dimensions: ***plíce*** [LUNGS] → ***ucho*** [EAR]; ***ledvina*** [KIDNEY] → ***žíla*** [VEIN]. Whereas words with the highest dissimilarity (< .15) were not as easy to categorise, but were still intuitively different, but not antonymic: ***babička*** [GRANDMOTHER] × ***fetovat*** [TO TAKE DRUGS]; ***striptér*** [STRIPPER] × ***Vánoce*** [CHRISTMAS].

At the individual word level, the cosine similarities can identify nearest neighbours for any target word in the data set, again with intuitive results. For example, inspecting the words closest to the target words in bold below demonstrates that this measurement can capture socio-semantic similarity:

školka [KINDERGARTEN] → *dupačky* [BABY ONESIE]; *přesnídávka* [BABY FOOD]; *pískoviště* [SANDBOX]; *houpačka* [PLAYGROUND SWING]; *dítě* [CHILD]

kostel [CHURCH] → *náboženský* [RELIGIOUS]; *náboženství* [RELIGION]; *věřící* [BELIEVER]; *modlit se* [TO PRAY]; *křeštanství* [CHRISTIANITY]

empatická [EMPATHETIC] → *soucitná* [COMPASSIONATE] *zdvořilá* [POLITE]; *spravedlivá* [FAIR]; *vyrovnaná* [BALANCED]; *vnímavá* [PERCEPTIVE]

tweetovat [TO TWEET] → *sociální síť* [SOCIAL NETWORK]; *blog* [BLOG]; *pop* [POP MUSIC]; *chat* [WEB CHAT]; *lajkovat* [TO LIKE]

Looking at the words that are most dissimilar from the targets also captures intuitive socio-semantic dissimilarities:

školka [KINDERGARTEN] × ***homofobie*** [HOMOPHOBIA]

kostel [CHURCH] × ***fet'ačka*** [FEMALE JUNKIE]

empatická [EMPATHETIC] × ***prezident*** [MALE PRESIDENT]

tweetovat [TO TWEET] × ***babička*** [GRANDMOTHER]

General Discussion

We have introduced the first large scale quantification of socio-semantic representations through 5 distinct dimensions of associative meaning (Study 1). This is, to our knowledge, the first data set that contains human derived ratings for how words are associated to 5 distinct dimensions of socio-semantic representation. We provide several aggregated variables, so that each word can be quantitatively represented with an associated normative value. We believe that the ratings provided here can easily be used for a range of different experimental applications or to explore how socio-semantic representations operate across the whole data set. For example, many studies have used measures of similarity for manipulation in memory experiments (Montefinese et al, 2015), evaluating models of semantic space (Hill et al. 2015) or for understanding statistical regularities in language (Dautriche et al., 2017).

We aimed to demonstrate the utility of the data from Study 1 by computing a measure of socio-semantic similarity, adopting a similar approach to Wingfield and Connell (2021). In Study 2 we used a high dimensional vector space containing proportion ratings across the different socio-semantic dimensions, from which words with similar (or dissimilar) socio-semantic profiles can be obtained by calculating the cosine distance between words. We believe that such a measurement will be of interest to researchers as it is derived specifically from theoretically meaningful dimensions of word representations, enabling researchers to select words that are proximally close or far apart in terms of the socio-semantic information they represent. However, a more rigorous evaluation of how well the distances capture human similarity judgements (see e.g. Verheyen et al., 2020) is clearly needed, which is a goal of future research.

We acknowledge that the depth of analyses presented in this paper does not address a number of open research questions. However, ongoing work that was outside of the main aims of the present paper is focused on further exploring a range of different topics. We are currently collecting data from a demographically more diverse population (see Preininger et al., 2022) in order to explore how age and gender might influence the socio-semantic representation of words (as has been done for other semantic dimensions, e.g. Warriner et al., 2013). We are also conducting a more detailed analysis of the role of linguistic variables (such as part of speech and GG) to further understand questions related to whether grammatically feminine words are rated differently to their grammatically masculine equivalents (Montefinese et al., 2019). In addition to exploring how word embedding models can be used to extrapolate ratings for a larger set of words (as suggested by Snefjella and Blank, 2020).

In summary, we hope that the work presented here highlights the potential for investigating socio-semantic dimensions of word meaning at scale, which can further our understanding of how words are represented and processed in the brain.

Supplementary Material

All data and code can be accessed at: https://osf.io/e47u8

Acknowledgements

This research was supported by a PRIMUS grant (PRIMUS/21/HUM/015) awarded to JB. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

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.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, *41*(8), 2149-2169.

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*(4), 1088-1097.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Krijthe, J. H. (2015). Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R package version 0.13, URL https://github. com/jkrijthe/Rtsne*.

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.

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*(3), 1271-1291.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological research*, *79*(5), 785-794.

Montefinese, M., Ambrosini, E., & Roivainen, E. (2019). No grammatical gender effect on affective ratings: evidence from Italian and German languages. *Cognition and Emotion*, *33*(4), 848-854.

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

Preininger, M., Brand, J., & Kříž, A. (2022). Exploring age and gender differences in socio-semantic representations of words. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*.

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., ... & Schwartz, H. A. (2014, October). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1146-1151).

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*(3), 1258-1270.

Snefjella, B., & Blank, I. (2020, September 24). Semantic Norm Extrapolation is a Missing Data Problem. https://doi.org/10.31234/osf.io/y2gav

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2021, August 3). Rating Norms Should be Calculated from Cumulative Link Mixed Effects Models. https://doi.org/10.31234/osf.io/3vgwk

Verheyen, S., White, A., & Storms, G. (2020). A comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for obtaining similarity data in the conceptual domain. *Multivariate Behavioral Research*, 1-28.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Wingfield, C., & Connell, L. (2021, September 20). Sensorimotor distance: A fully grounded measure of semantic similarity for 800 million concept pairs. https://doi.org/10.31234/osf.io/fq53w

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic?: Iconicity in English sensory words. *Interaction Studies*, *18*(3), 443-464.