SocioLex: Quantifying the socio-semantics of Czech

Mikuláš Preininger, James Brand, Adam Kříž (Charles University)
mikulas.preininger@ff.cuni.cz

The ability to form rich semantic representations for thousands of words is a defining feature of human cognition. In order to better understand these representations and the way they contribute to lexical processing, researchers have applied various methodologies that that quantify the different ways we associate semantic information with individual words in the form of large scale norming databases [1, 2]. We extend this line of research by investigating variables that capture socially meaningful dimensions of semantic representation, in order to map out the socio-semantics of thousands of Czech words. We present ongoing research that aims to collect normative ratings for nouns, adjectives and verbs, along five theoretically distinct dimensions of social meaning. Czech has been relatively understudied in terms of the availability of normative ratings, with current datasets limited in terms of size and scope [3, 4, 5]. Moreover, in the past 30 years the population has experienced dramatic social shifts, making it an ideal candidate to study differences and similarities in semantic representations and how this affects language processing across populations with varying social profiles.

Our study comprises a word list of 2,700 Czech words, which captures a broad range of semantically diverse domains (such as occupations, religion, nature), with varying lexical frequencies ([6]) and grammatical gender variants (see supplementary page for further details on the morphology of Czech). Participants were asked to rate lists of 105 words along each of the five dimensions: *gender*, *political alignment*, *location*, *valence* and *age*. All scales (except for the age dimension) were anchored by two contrasting parameters, along a 7-point Likert scale, each with a neutral midpoint option. For *gender* the scale was anchored by masculinity/femininity, for *political alignment* the anchors were liberalism/conservatism, for *location* it was ruralness/urbanness and for *valence* it was positive/negative. For the *age* variable, participants could choose from age categories of 0-7, 7-17, 18-30, 31-50, 51-65, 66-80 and 81+ years old and were allowed to choose multiple options. We have currently recruited 256 Czech speakers (49 males) with an age range of 19-40 (mean = 22.8, SD = 4.8), who also completed a basic socio-demographic profile questionnaire, whereby they self-reported on their gender, political alignment, location and optimistic/pessimistic character, using similar scales as used for the word ratings. The distribution of participants in terms of their socio-demographic profile is shown in Figure 1.

An initial inspection of the data indicates that participants encode social meaning in distinct ways. Figure 2 provides a visualisation of the data for two different words, *bohyně* (Goddess) and *bůh* (God), we can see that although these words are semantically very similar to one another, our socio-semantic dimensions capture subtle differences in their individual meanings. In order to better understand how these variables operate across the full list of words and for different parts of speech categories, we provide an interactive app that allows users to explore correlations between the variables, in addition to visualising the dataset in 2 dimensional space through t-SNE plots, where words with similar ratings appear closer to each other.

The app can be accessed at: https://jamesbrandscience.shinyapps.io/sociolex/

Whilst these initial insights provide a starting point for understanding the way socio-semantic variables operate, as Fig. 1 shows, we currently have a very biased sample of participants in terms of their social profiles, but as data collection progresses (beyond our current sample), we hope to use this information to better understand where differences and similarities exist between participants from different socio-demographic backgrounds, thus providing a more diverse and textured database for which we can better understand how social meaning is encoded and processed, not only in a large sample of semantically diverse words, but critically, for a large sample of diverse participants.

References

[1] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior research methods*, pp. 904–911, 2014.

[2] D. Lynott *et al.*, "The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words," *Behavior Research Methods*, pp. 1–21, 2019.

[3] J. Misersky *et al.*, "Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak," *Behavior Research Methods*, pp. 841–871, 2014.

[4] M. Luniewska *et al.*, "Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian," *PloS one*, e0220611, 2019.

[5] A. Kříž and F. Smolík, "Ratings of imageability, concreteness, specificity, familiarity and age of acquisition in Czech nouns and verbs: Their mutual relations and dependencies," *Československá psychologie*, p. 507, 2015.

[6] M. Křen, "Czech National Corpus in 2020: Recent Developments and Future Outlook," *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pp. 52–57, 2020.

Figures

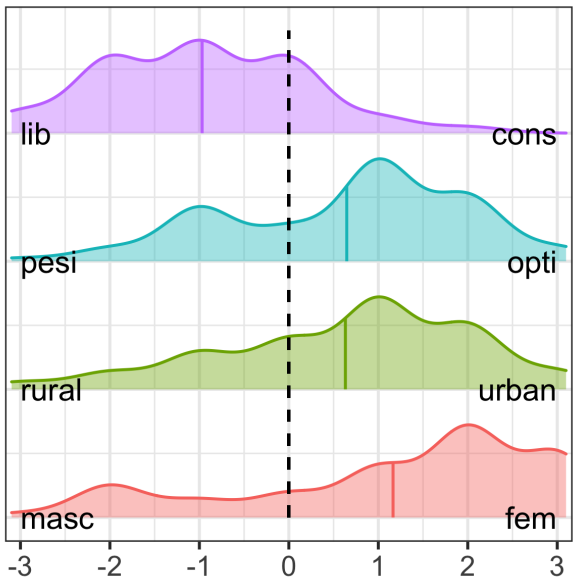

Figure 1: Density plot of the socio-demographic profiles of participants (n = 261) along four variables (from top to bottom) *political alignment* (liberal-conservative), *pessimist-optimist*, *location* (rural-urban) and *gender* (masculine-feminine). The x axis represents a numeric transformation of the Likert scale values, with larger absolute values representing a response that is equivalent to *very*, e.g. -3 for the political alignment variable represents *very liberal*. The dashed line at 0 represents a neutral midpoint option. The solid verticals lines indicate the mean values.

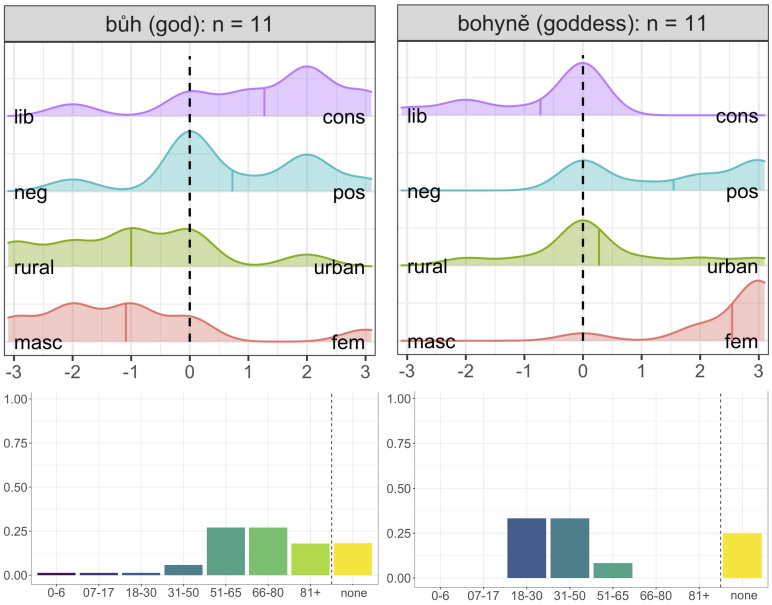

Figure 2: Visualisation of ratings along the five dimensions for the words *bůh* (god) and *bohyně* (goddess). The density plots show the distributions of ratings along four dimensions (from top to bottom): political alignment (liberal-conservative), valence (negative-positive), location (rural-urban) and gender (masculine-feminine). The dashed line indicates a neutral midpoint option. The bottom two barplots capture the distribution of ratings along the age dimension, with the y-axis representing the proportion of response (the yellow bar *none* indicates an option for no age association).

Supplementary information

Grammatical gender in Czech

Due to the combination of facts that one of the explored dimensions is gender and that Czech has a category of grammatical gender (GG), a remark on GG in Czech and the consequences for our research will be useful.

In Czech, there are four values of GG: feminine, maculine animate, masculine inanimate, and neutral. GG works differently for different parts of speech. For verbs and adjectives, GG is a so called syntagmatic category – GG of verbs and adjectives is not inherent but is determined by the grammatical gender of the NP/Pron the verb or adjective relates to. On the other hand, GG in nouns is a so called paradigmatic category which means that every noun has its inherent value of GG. GG of nouns is motivated by several factors. If the noun refers to an animate object, its GG tends to be consistent with the natural gender (i.e. sex) of that object. One exception from this tendency is the presence of generic masculine in Czech: when a speaker refers to one or more objects and cannot be sure of their sex, he or she is likely to use the masculine variant. In cases when the noun refers to an inanimate object, its GG is motivated formally, by the stem morpheme (however, due to the changes over time this formal motivation is not always transparent).

Given the fact that grammatical gender and natural gender on the one hand and gender as a social category on the other hand are separate, yet tightly related phenomena, we decided to design the wordlist in a way that would enable us to control for this relationship as much as possible. To achieve that we treat grammatical and natural gender as predictor variables and associated gender as an outcome variable. This does not concern verbs since in the questionnaire they are presented in infinitive forms and thus do not express GG. Nouns and adjectives, on the other hand, express GG since they are presented in nominative singular. We decided to manipulate the grammatical and natural gender whenever it was possible. For most adjectives, we included masculine as well as feminine forms (e.g. dobrý ‚good' [masc.] + dobrá ‚good' [fem.]). For nouns referring to animate objects, we included masculine as well as feminine variants (strýc 'uncle' as well as teta 'aunt' or pekař 'male baker' + pekařka 'female baker').

Recent historical development and Czech

In the past several decades, the Czech Republic has undergone a number of significant shifts (e.g. Soviet occupation in the 1968, Velvet revolution in the 1989, entering EU in the 2004 etc.). This is especially important in relation with the following two facts. First, lexicon, as opposed to grammar, is a domain of language that is highly sensitive to changes in the objective reality (e.g. emergence of new entities, changes in linguistic situation etc.). Second, semantic representations are modulated mainly by our experience. This combination indicates that exploring Czech from the presented perspective may uncover various and unique patterns speakers conceptualize the world around them.