The Relation of Categories of Concreteness and Specificity: Russian Data

Ivanov V.V.
Innopolis University
Solovyev V.D.
Kazan Feredal University

Concreteness vs Specificity

диван (sofa) – мебель (furniture)

- sofa is a more concrete concept and at the same time more specific
- furniture is more abstract and more general
- The main research question: are concreteness and specificity substantially different categories or closely related (highly correlated)?

Definitions

- Concrete concepts are those that are perceived by the senses. The quantitative measure of specificity is indicated in specially created dictionaries.

- Specificity is characterized by the places of concepts in a hierarchically ordered structure of concepts. The closer to the leaves, the more specific.

Dictionaries with indices of specificity

- Created by surveys of native speakers
- English. Dictionary for 40 thousand words
- (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)
- Russian. Dictionary for 1 thousand words
- (Solovyev V. D., Ivanov V. V. and Akhtiamov R. B.: Dictionary of Abstract and Concrete Words of the Russian Language: A Methodology for Creation and Application. Journal of Research in Applied Linguistics. vol. 10, 215 - 227, 2019)
- Machine extrapolation of respondents' estimates. Machine dictionary for the Russian language, 22 thousand words

  (https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html)

Concreteness ratings for the Russian language

- Respondents are students of Kazan Federal University and Belarusian State Pedagogical University
- At least 40 ratings on a 5-point scale for each word
- 1000 most frequent nouns according to Lyashevskaya-Sharov's dictionary (2009)
- Free access https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html
- The main disadvantage is the small size

Machine dictionaries

- Main idea:

- build a vector model of words for a large corpus of texts
- apply neural networks, classifiers to extrapolate ratings based on distances in the semantic word space

- Evaluation of the quality of extrapolation is carried out by comparison with human ratings (Spearman correlation)

- Best achieved score for English - 0.9

Machine dictionaries for the Russian language

- The first version is a dictionary of word forms (not lemmas). It contains 88 thousand word forms (nouns and adjectives) and is based on the Google Books Ngram corpus. When it was created, an original method was implemented, based on the idea that concrete words are often found in texts together with concrete ones, and abstract ones – together with abstract ones.
- The second version of the computer dictionary was created using the word2vec technology, the fastText model. It contains 64 thousand words (lemmas).
- The third version is a 22,000-word machine dictionary built using deep learning technology, the BERT model.
- The correlation coefficient between the human ratings and the third variant of the machine ratings is 0.81 according to Spearman.

Assessment of the degree of specificity

- Thesauri WordNet and RuThes are used

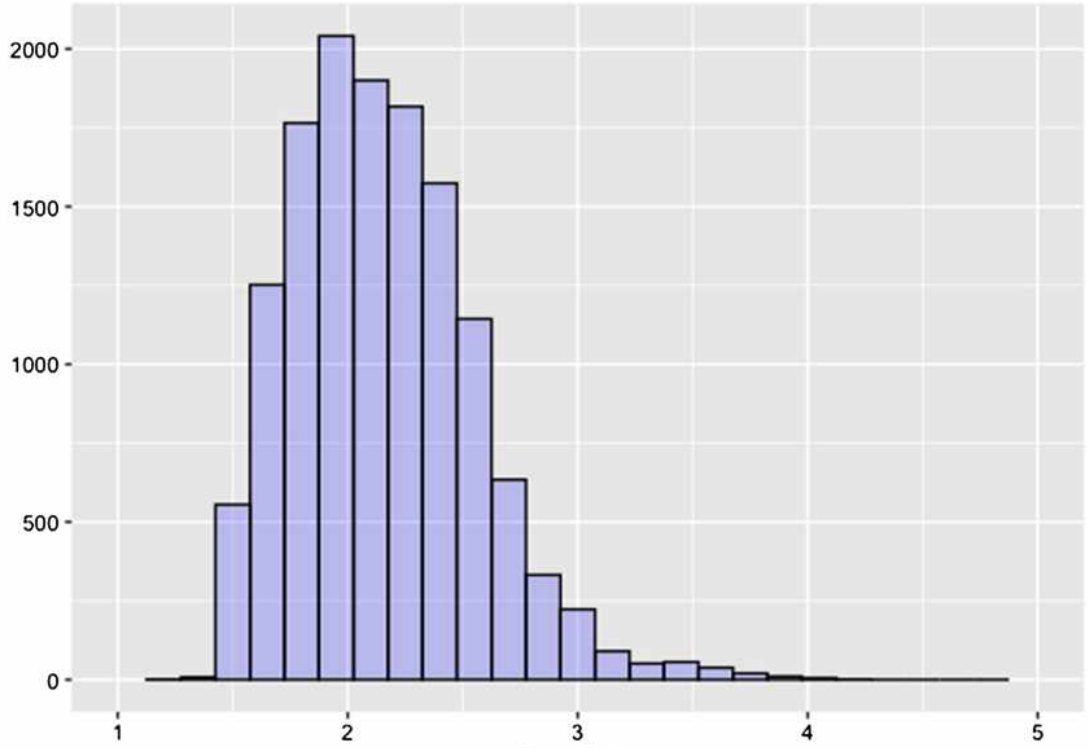
- = (1 + d) ⁄ D, where d is the total number of hyperonyms of the target word and D is the maximum distance from leaves to the top of the hierarchy. For WordNet D = 20, for RuThes - D = 13.
- (Bolognesi M., Burgers Ch., Caselli T. On abstraction: decoupling conceptual concreteness and categorical specificity. Cognitive Processing (2020) 21:365–381)

- Other formulas were considered that lead to similar results

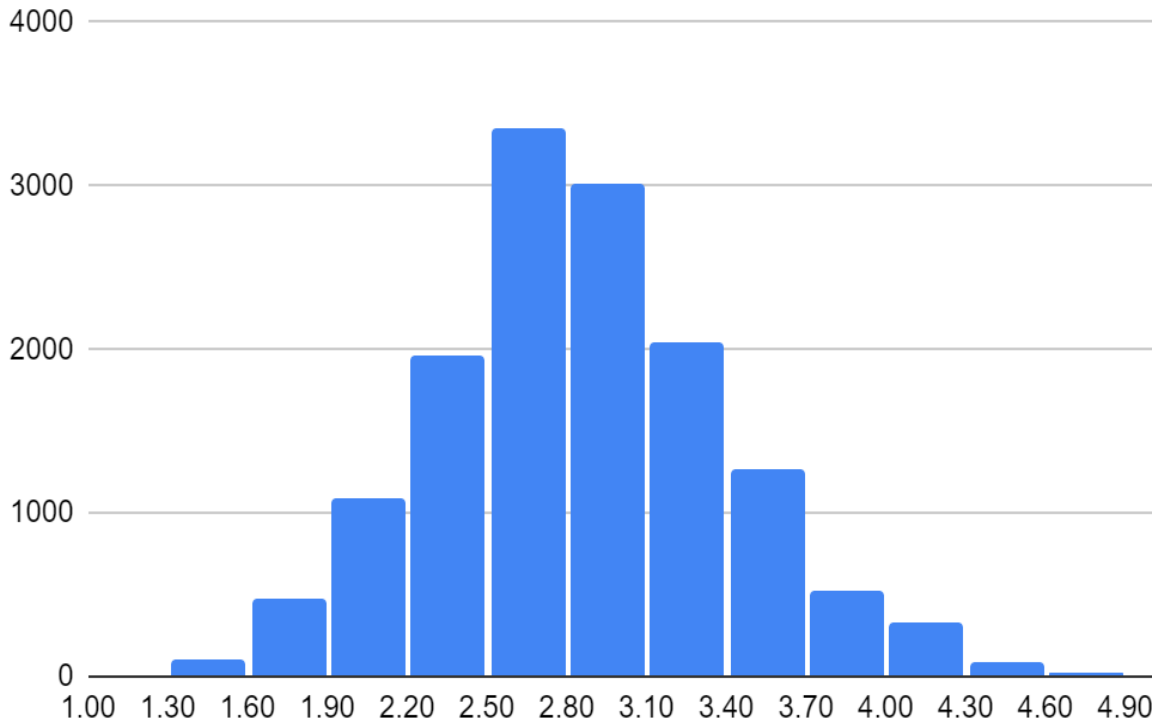
Correlation of indices of concreteness and specificity

- For the Russian language, Spearman's correlation coefficient = 0.264, Pearson = 0.256 ($p < 0.001$).
- For English, the coefficients are 0.361 and 0.354, respectively.
- Thus, the degree of correlation between these concepts is low.

Comparison of distributions of specificity ratings for English and Russian



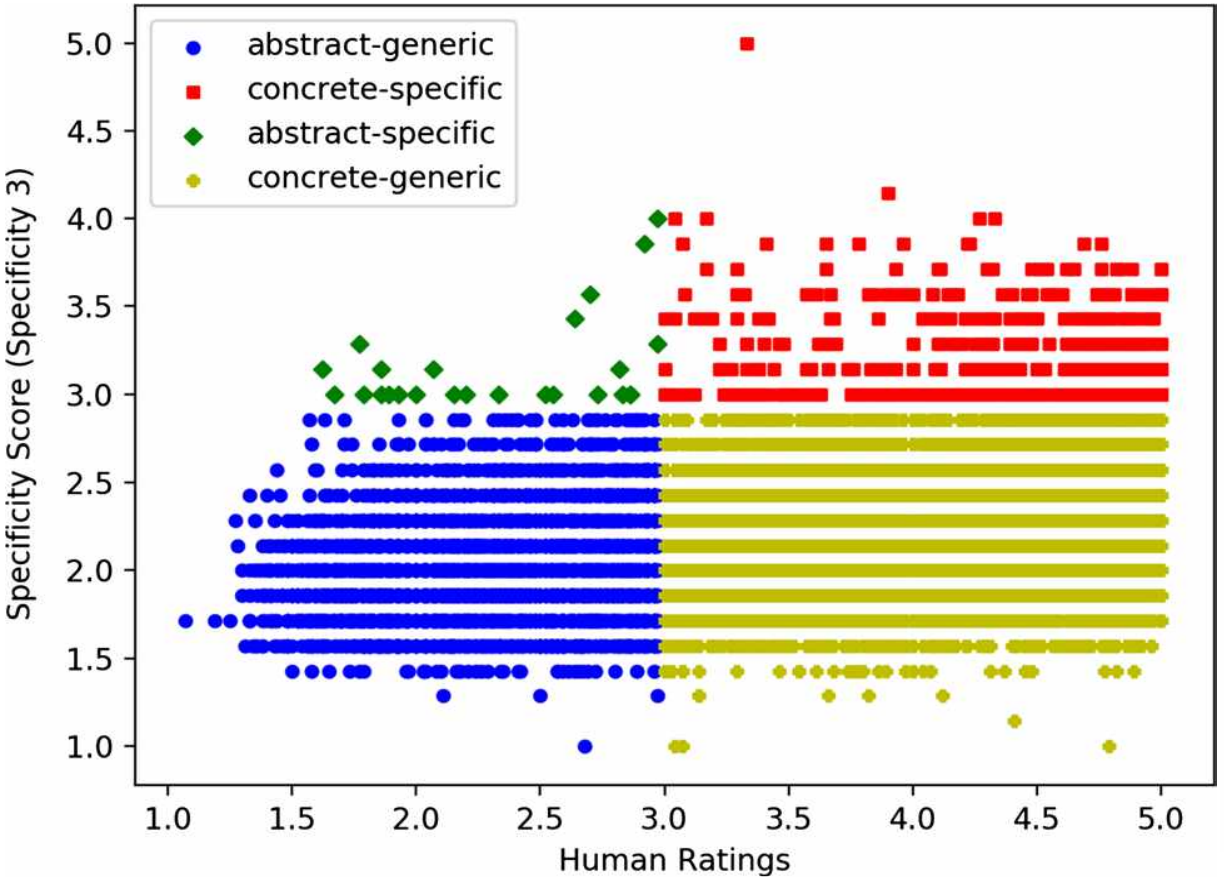
English



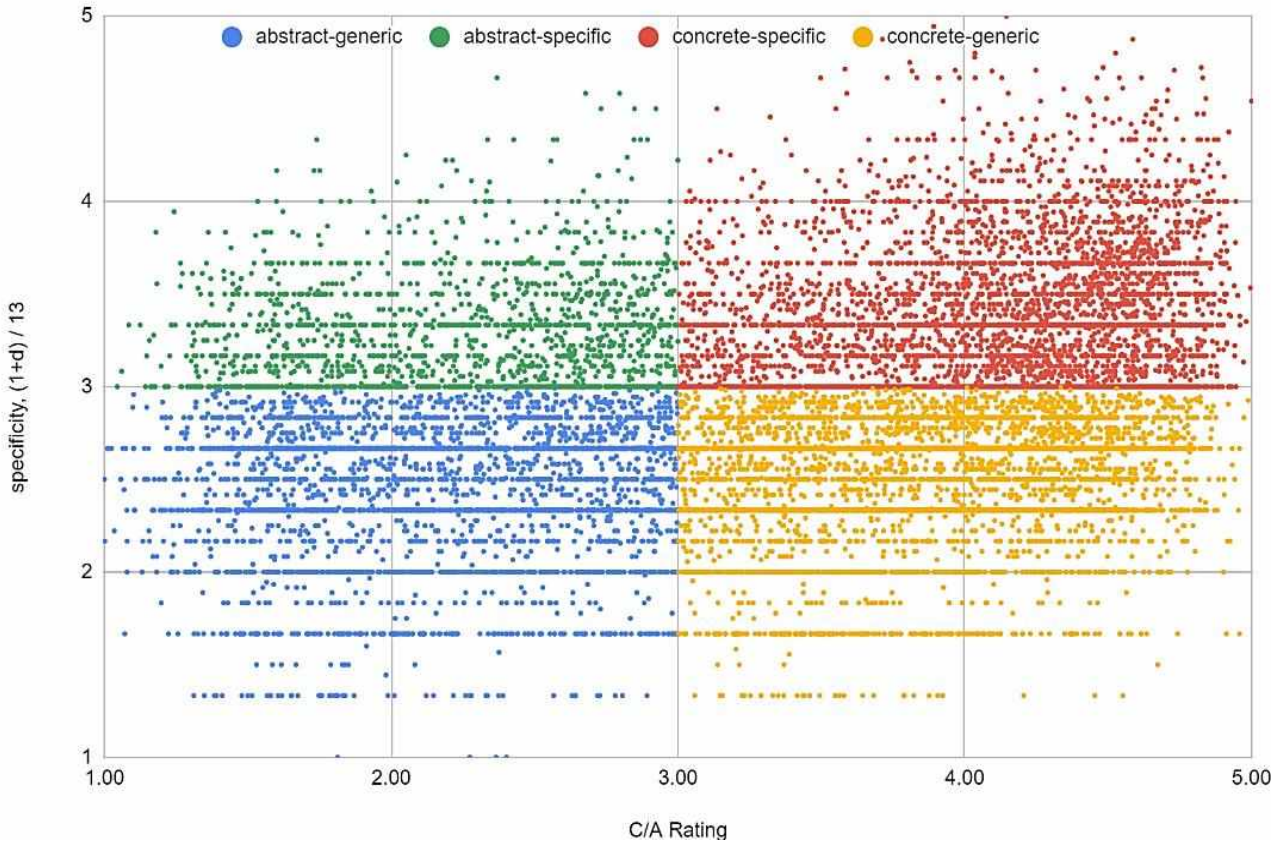
Russian

5 – highest specificity

Distribution of words in two-dimensional space concreteness-specificity



English



Russian

Words with extreme ratings

- Upper Right Quadrant: highly specific and highly concrete

- in English, words with extreme meanings are: *karaoke*, *epinephrine*, *aspirin*, *heifer*, *triglyceride*, *glucose*, *chloroform*, *fructose*, and *petroleum*

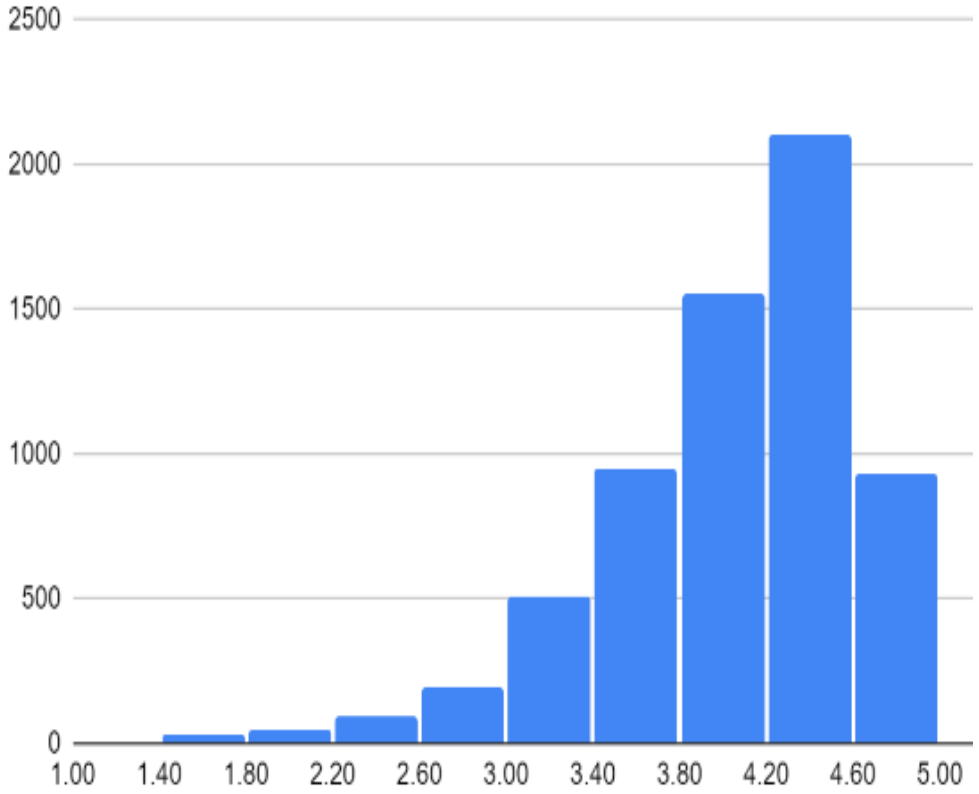
- in Russian: *травмпункт, мегафон, бомбардировщик, психбольницу, радиотелефон, мобильник, домофон, монитор, госпиталь, ноутбук, больница, горбольница, медпункт, эвакуатор, холера*

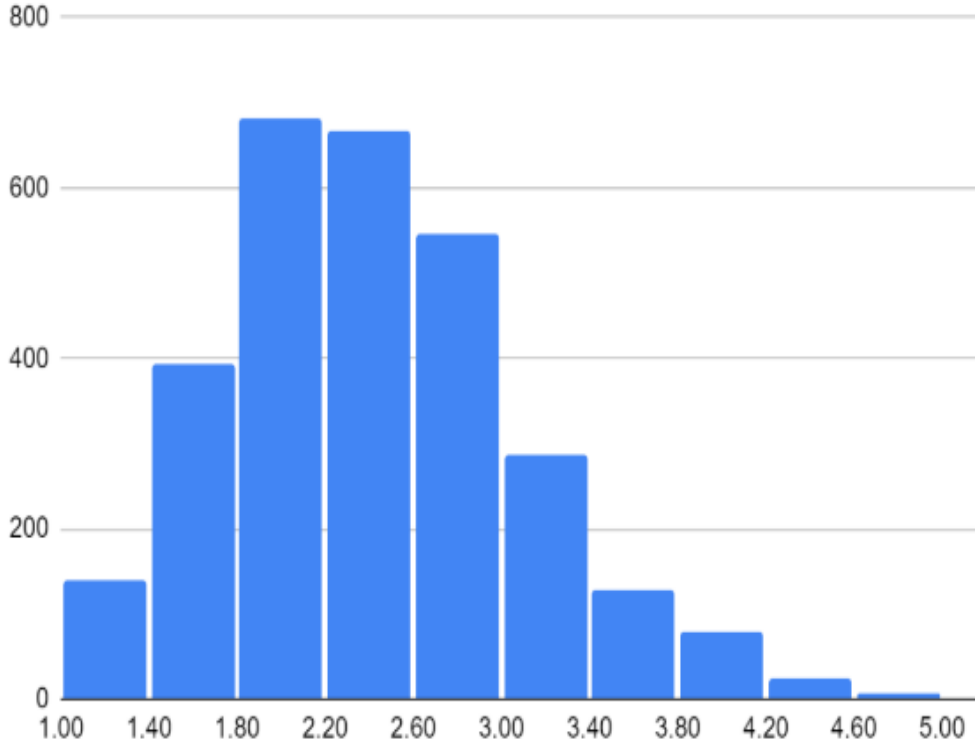
- Upper Left Quadrant : highly specific and highly abstract

- in English, this sector presents: *cakewalk, fundamentalism*, and *vintage, bootleg*, *finisher*, *general*, *mankind*, *monotheism*, *polytheism*, and *summons*

- in Russian: *идолопоклонство, кощунство, поругание, святотатство, плодородие, сретение, роскошество, помрачение, заикание, роскошь, царствование*

Concreteness indices for hyponyms of nodes PHYSICAL ESSENCE and ABSTRACT ESSENCE





Conclusion

- It is shown that concreteness and specificity are essentially different concepts. The Russian language data confirmed the English data.
- RuThes is more balanced than WordNet.
- The division of concepts into abstract entities and physical entities, presented in RuThes, is consistent with indices of concreteness.

Thank you for attention!