Supplementary notes for *Size Matters: towards a critical assessment of size-related effects in gender assignment*

Contents:

Note 1. Absolute size datasets

Essentially, both Absolute Size Hypothesis and Categorial Size Hypothesis are tested in the same way: we use experimental design to arrange concepts on a scale of size, process the resulting datasets, map concepts to the lexical items of our target languages and test if Gender 3 and Gender 4 values are distributed along the scales in a way which is non-random; more specifically that Gender 3 nouns tend towards the *bigger* end of the scale, and Gender 4 nouns tend towards the *smaller* end of the scale.

1. Compilation of the absolute size datasets

Absolute Size Hypothesis in this study was formulated as follows:

Absolute Size Hypothesis: *Is it true that, across lexicon, there is a tendency to assign smaller entities to Gender 4 and/or bigger entities to Gender 3?*

The hypothesis was tested independently on several datasets of concepts annotated for size. Two datasets come from the existing semantic feature databases (Mcrae et al. 2005, Binder et al. 2016). In these, for hundreds of different objects, their featural representation is experimentally collected. Detailed information on how exactly the feature values of concepts in the two 'external' datasets were calculated, as well as the links to the complete datasets, are to be found in the respective publications.

As an alternative to McRae et al. (2005) and Binder et al. (2016), we considered using a more recent and much larger (as compared to) semantic feature database by Buchanan et al. (2019) which contains feature representation for around 4,000 concepts. However, this collection contains results of the studies where semantic features assigned to nouns are essentially extractions from complex definitions/descriptions of concepts originally provided by subjects. For example, a concept of “department” has a feature ‘large’ assigned to it several times but at least some of these assignments come from the multi-word descriptions of *department* such as ‘a small part of a larger entity’ where *large* has nothing to do with a prototypical size of a department. Due to this we decided to rely on two smaller datasets by McRae et al. (2005) and Binder et al. (2016) which only include ‘small’ and ‘big’ as more straightforwardly assigned size values.

The third dataset, Dataset Macabre, was compiled in an online experiment we ran ourselves. Fig. 1 below shows the web-interface we used. The experimental task was formulated as follows (translated from Russian): “Please, compile a list of large (in the other part of the same experiment — small) objects, with each word separated by a comma. Do not use multiword expressions and derivations from other words (authors: that meant avoid diminutive and augmentative derivations). You have ten minutes to compile the list, after that your answer will be automatically stored. You can send your form by clicking the ‘continue’ button if you are done before the time is over”. The query for big and the query for small objects were presented to different subjects in random order. The online version of the experiment is available at https://sizematters.pythonanywhere.com/.

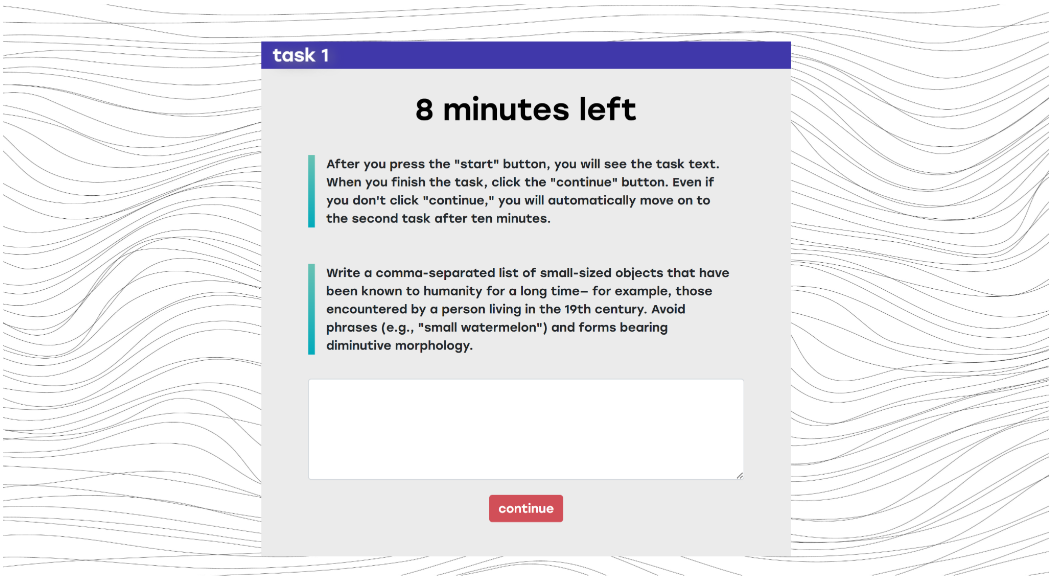

Figure 1. Interface of the absolute size experiment

The general characteristics of the datasets are shown on the following figures:

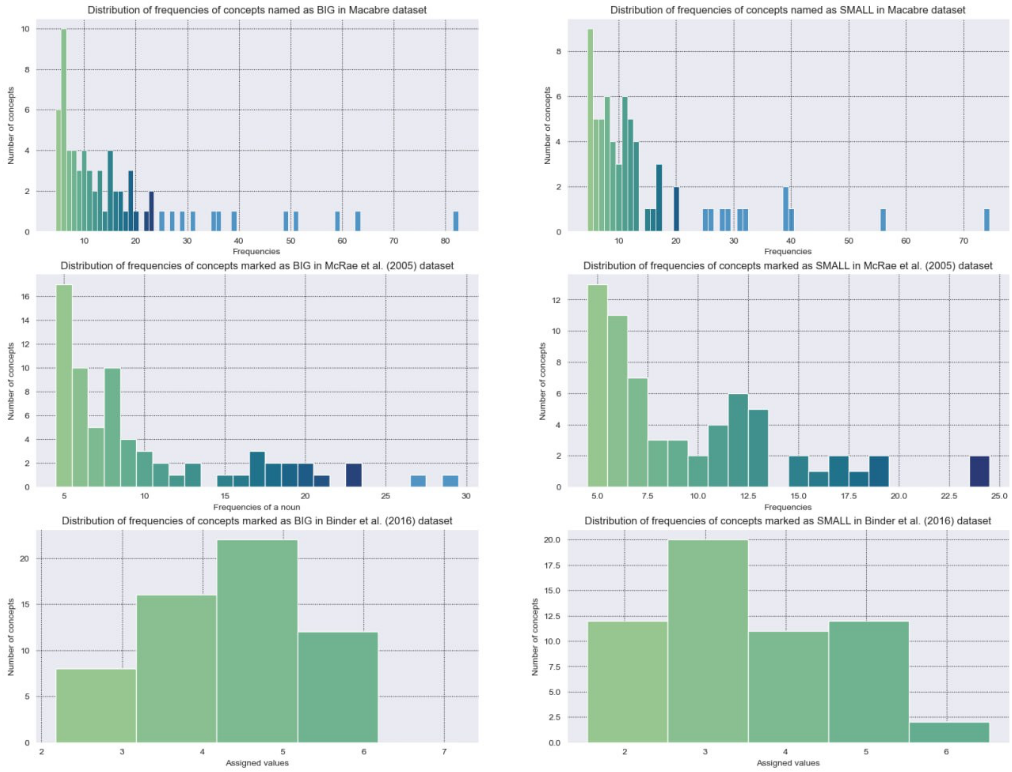

Figure 2. Distribution of different size values in three datasets.

The figure shows density plots of distribution of concepts on the scale of size in each dataset. For McRae et al. (2005) and the Dataset Macabre, the size value of a concept (X-axis) is simply the number of subjects who, respectively, indicated size as a salient property of the concept (McRae et al. 2005) or produced the name of the concept when asked to give an example of a big or small entity (Dataset Macabre). To be included, a concept should have been named at least 5 times (a filter the Dataset Macabre inherited from McRae et al. 2005, for which the raw list containing the concepts named less than five times is simply unavailable). As a result, for McRae et al. (2005) and for the Dataset Macabre, the minimal frequency value on the charts is five. For Binder et al. (2016) dataset, size values are distributed from six to zero. In this experiment, subjects were asked to estimate to what degree — on the scale from zero to six — a particular concept from a fixed list (like McRae et al. 2005) is associated with a particular semantic characteristic, also from a fixed list of characteristics (unlike McRae et al. 2005), including size properties.

2. Data filters applied to the absolute size datasets

All three experimental datasets originally featured some concepts that we did not include in the final datasets.

In the case of Binder et al. (2016), we removed the words belonging to parts of speech other than nouns. In the case of the Dataset Macabre, we removed the answers that failed to follow initial restrictions (no singulatives, no diminutives or augmentatives, no multiword expressions).

Further, from all three datasets we removed human nouns, because these cannot be assigned to Gender 3 or Gender 4 in East Caucasian (see Section 2 of the paper). We also removed words that, in our source languages English and Russian, may refer to things other than concrete objects (i.e. abstract and mass nouns) with clear limits in space and measurable dimensions. We did so because, for such nouns, it is not altogether clear what the assigned size value stands for. For example, in the dataset by McRae et al. *rice* is characterized as small by 12 subjects out of 30. Similarly, in the Dataset Macabre for small we have 6 entries for *pesok* 'sand'. These substances are collections of particles constituting their inner structure. With reference to rice and sand, 'small' rather means 'fine-grained'. Neither English, nor Russian word is naturally interpretable as a name for a bounded real-world entity that is a subject to size evaluation, i.e. *pesok* is not used to denote one grain of sand. We thus can reasonably doubt that the size score 6 obtained by sand in our experiment should be interpreted as the subject's characterization of the concept denoted by the word *pesok* 'sand'. Moreover, the languages of the study also lack singulative lexicalizations. As a result, the reduced datasets only contained names for entities to which size property is applicable in an unequivocal way.

Figure 3 represents the process of annotating concepts for size in the absolute size experiment.

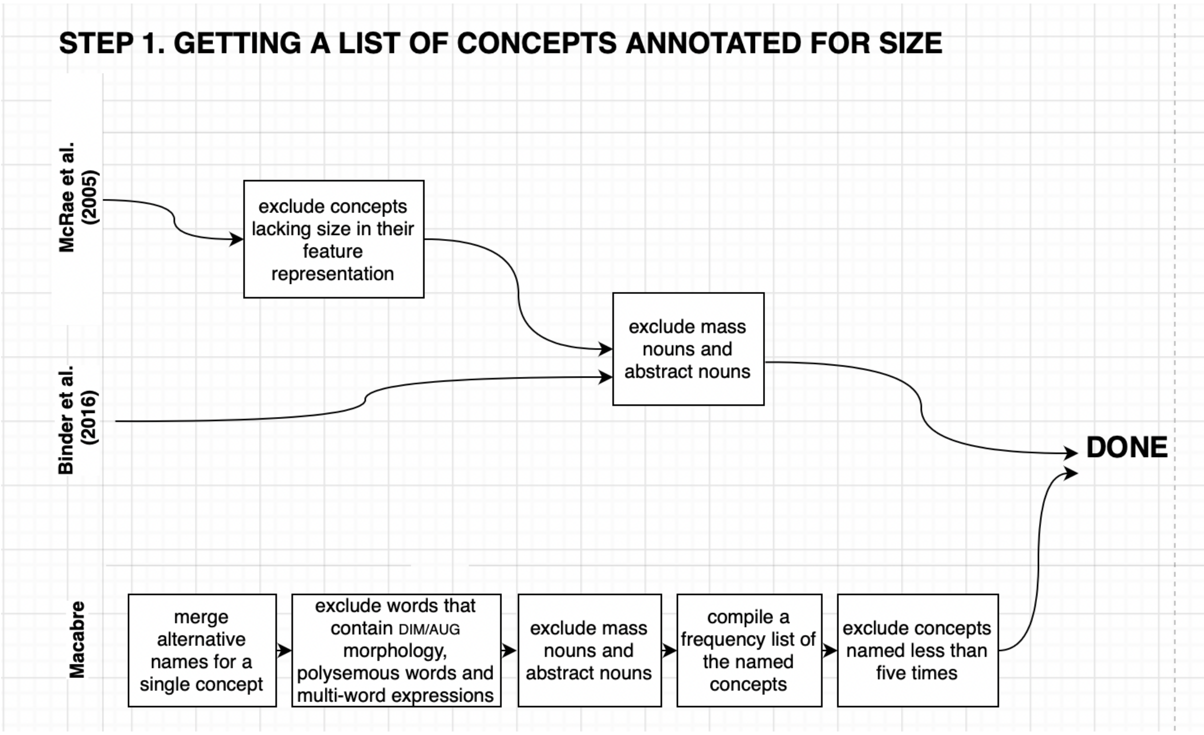

Figure 3. Getting a list of concepts annotated for size: the workflow

Note 2. Comparison of the absolute size datasets

All concepts used in testing the Absolute Size Hypothesis are available on GitHub. Section 4.1.4 of the paper briefly compares the datasets in terms of their content. Below, we provide a few more details. Figure 4 shows the similarity between the three datasets in terms of the number of concepts they share, small on the left, big on the right. Thus, all three datasets share 4 concepts for small, while McRae et al. (2005) and Binder et al. (2016) share 9 concepts for big that are not shared with Dataset Macabre.

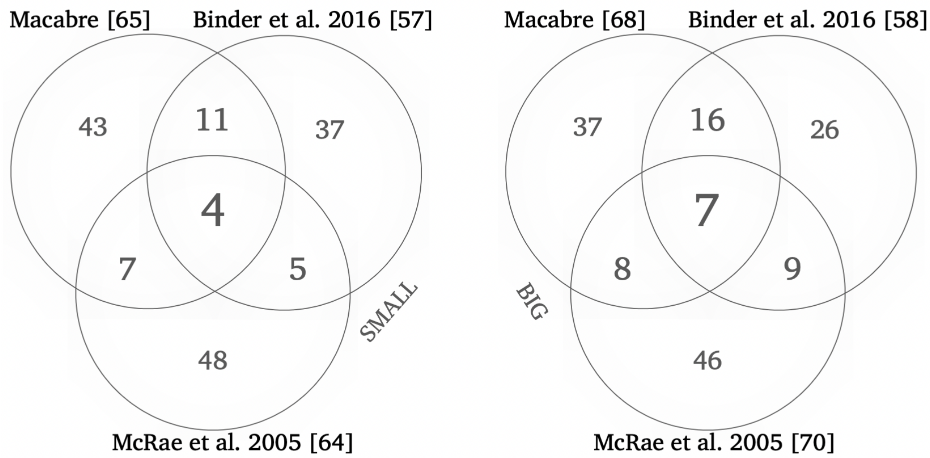

Figure 4. Overlap between three datasets

It is clear from the figure that Dataset Macabre and Binder et al. (2016) are more similar to each other than any of these to McRae et al. 2005, both for small and big; and that datasets for big are generally more similar to each other than the datasets for small.

Yet, all in all the datasets are very different, and even Dataset Macabre and Binder et al. (2016) share less than one third of the list for big. For specific lexical items shared by the datasets, cf. Table 1 (small) and Table 2 (big).

concepts	datasets	number
ant, key, mouse, stone	all three	4
butterfly, comb, flute, hamster, penguin	mcrae + binder	5
beetle, bullet, coin, flea, fly, pin, thimble	ma&r + mcrae	7
bird, book, egg, eye, feather, finger, flower, hairbrush, mosquito, pen, pencil,	ma&r + binder	11

Table 1. Concepts for small entities present in more than one dataset

concepts	datasets	number
bridge, church, elephant, horse, oak, rocket, whale	all three	7
bus, camel, crocodile, house, limousine, piano, submarine, tiger, truck,	mcrae + binder	9
barrel, bear, building, cannon, cart, house, ship, skyscraper	ma&r + mcrae	8
bed, car, carriage, cloud, field, forest, island, lake, mountain, plane, river, sun, table, train, tree, volcano	ma&r + binder	16

Table 2. Concepts for big entities present in more than one dataset

Datasets align concepts along the size of scale. But even the concepts on the top of the lists are also very different, especially in the lists for small; cf. Table 3 below (which is a copy of Table 5 in the paper).

a. small			b. big		
Macabre **[65]**	McRae et al. (2005) **[64]**	Binder et al. (2016) [**57**]	Macabre **[68]**	McRae et al. (2005) [**70**]	Binder et al. (2016) [**58**]
needle 74	flea 24	ant 5,87	mountain 82	elephant 29	whale 5,83
ring 56	hamster 24	mosquito 5,7	elephant 63	whale 27	mountain 5,76
button 39	*chickadee* 19	mouse 5,43	ship 59	bear 23	*sun* 5,74
coin 39	*cottage* 19	butterfly 5,13	house 51	gorilla 23	volcano 5,68

earring 39	ant 18	*bee* 5,1	*ocean* 49	truck 21	rocket 5,6
mouse 32	*mole* 17	hamster 5,03	*sea* 39	*buffalo* 20	*airport* 5,86
feather 31	mouse 17	egg 4,93	*castle* 36	*lion* 20	plane 5,5
insect 28	pin 16	*toe* 4,93	*tower* 35	horse 19	elephant 5,41
drop 28	*canary* 15	*ticket* 4,82	whale 31	*yacht* 19	bridge 5,38
spoon 26	*sparrow* 15	*cellphone* 4,74	*planet* 29	*bull* 18	*zoo* 5,37

Table 3. Top-10 rankings from 3 studies

In Table 3, the italicized concepts are those top concepts for small and big that are not members of the other two list. For Dataset Macabre and McRae et al. (2005), the number next to the name of the concept indicates the number of subjects who named it as small or big; and for Binder et al. (2016) the number is the average evaluation of size obtained in this experiment. Only 'mouse' and 'feather' from Macabre top-ten are present in at least one of the other two datasets for small; and 'feather' did not make it to the Binder top-ten, the list where it is present. Once again, the datasets for big seem more comparable.
We assume the reason for these differences is the experimental design, with controlled concept lists for evaluation in McRae et al. (2005) and Binder et al. (2016) and a free-naming task in the case of the Dataset Macabre. Concepts for McRae et al. (2005) and Binder et al. (2016) were not preselected for size-relevance, so that very small and very big entities are not necessarily present in these two datasets; and while the top of the Dataset Macabre does include entities which are saliently small and big, but the set of concepts included quickly becomes random when going down, because of the nature of the free naming experiment.

One thing that we expect that IF a concept that is on the top of the Dataset Macabre also happens to be included in one or both of the two other datasets, THEN it also includes size features as a prominent component of its feature representation (McRae et al.) or the relevance of size is rated highly (Binder et al. 2016). This seems indeed to be the case: among the small, 'mouse', shared by all three datasets, is high in Macabre and present in two other lists, where it is also high; and 'feather', shared by Macabre and Binder et al. (2016) and immediately following 'mouse' in Macabre, is relatively high (16th position, 4,34) in Binder et al. (2016). Among the big, 'whale' and 'elephant', shared by all three datasets and among the top ten in Macabre, are both among the highest in the other two lists, etc. But 'needle' and 'ring', two smallest concepts in the Dataset Macabre, are simply not included in the other two datasets. If they were, they would probably make it to their tops.

Note 3. Aggregating individual rankings for categorial size datasets

Categorial Size Hypothesis in the study is as follows:

Categorial Size Hypothesis: *Is it true that, for animals and birds, there is a tendency to assign smaller species to Gender 4 and/or bigger entities to Gender 3?*

To test for this hypothesis we carried out an online experiment for size annotation (also described in Section 4.2 of the main article. We selected four categories that are, we assume, relatively easy for size evaluation, including animals and birds (i.e. the two categories where we expected to find a correlation between gender and size at least in Archi) and utensils and body parts (where no such correlation has been reported in previous literature). The design of the online experiment is accessible here: http://drakolit.pythonanywhere.com/.

Prior to computing the aggregated ranking, we conducted an assessment of inter-annotator agreement using Kendall's coefficient of concordance (W). Our subjects were allowed to assign the same rank to two different concepts, or not to evaluate a difficult concept altogether. To address ties and missing values, we utilized the implementation for randomly incomplete datasets from the irrNA package in R[1]. The outcomes revealed a significant level of agreement among our subjects (0.91, 0.82, 0.88, and 0.89 for animals, birds, utensils, and body parts, respectively), proving the viability of ranking aggregation.

Because of ties and missing values, our subjects' rankings could not be compared directly. For ties, we applied the following technique. The number of ranks was scaled up to the total number of concepts evaluated, and the concepts with a tie were assigned an average. Thus, the original ranking of {1, 2, 2, 2, 3, 4, 4, 5} would be first scaled up to 8 ranks {1, (2 = 3 = 4), 5, (6 = 7), 8} with the consequent averaging of ties {1, 3, 3, 3, 5, 6.5, 6.5, 8} We used the `rank` function from base R (R Core Team 2021).

Because respondents had an option to skip evaluation of concepts they deemed difficult, the total number of ranks varied across answers (even after scaling up described above). We needed to project all rankings onto a single scale. To address this, we implemented the MinMaxNormalization formula: $rank_i = \frac{X_i - min(X)}{max(X) - min(X)}$, where X represents the list of all

[1] https://cran.r-project.org/package=irrNA

ranks for a specific respondent, and i denotes the position of the value undergoing normalization. As a result, we scaled values to a standardized interval between zero and one. For the original {1, 2, 2, 2, 3, 4, 4, 5} example, the resulting normalized ranking is {0, 0.29, 0.29, 0.29, 0.57, 0.79, 0.79, 1}.

An alternative approach to inter-respondent rank normalization was also used. This involved computing a rank using the formula: $rank_i = \frac{x_i}{n+1}$, where n denotes the number of non-null values (i.e. the number of concepts ranked by the subject). This particular ranking methodology is designed to retain a greater number of ties and is not geared towards assigning each value to a distinct group. This calculation delivers ranks that are already normalized to 0 to 1 scale. For the above example of the original {1, 2, 2, 2, 3, 4, 4, 5} ranking, the resulting normalized ranking is {0.11, 0.33, 0.33, 0.33, 0.56, 0.72, 0.72, 0,89}.

For both methods of normalization, to generate an aggregated ranking, we computed the median of all rankings for each individual element. We computed Kendall's rank correlation coefficient (tau) to compare the outcomes of the two alternative methods of rank aggregation. The analysis demonstrated a robust correlation, with tau values surpassing 0.97 for all datasets. This suggests a high level of consistency between the two aggregation techniques. As an example of the two alternative rankings of one category, consider Table 4:

sparrow	sparrow
bullfinch	bullfinch & bat
bat	
swift	swift & swallow
swallow	
dove	
jackdaw	dove & jackdaw & cuckoo
cuckoo	
magpie	magpie & quail
quail	
crow	crow
crow	partridge
hen	hen
owl	owl & rooster
rooster	

eagle	eagle
crane & heron	crane & heron & stork
stork	

Table 4. Two alternative rankings of the *birds* category

The table does not show a single difference except that some concepts ranked with respect to each other in the first method of rank calculation are merged with a tie in the second method. As the final aggregated ranking we used the one obtained by the first technique, because it yields less ties and thus delivers more robust Mann-Whitney results. It is shown below (ties are marked with ampersands). For the online experiment, we included several insects into the category of animals; they were not included in the calculation of Mann-Whitney tests but served as a clear point of reference for being small; while 'elephant' was assumed to provide a similar clear point of reference for being big.

Animals

1. ant
2. fly
3. bee
4. spider
5. butterfly
6. frog
7. lizard
8. mouse
9. bat
10. mole
11. squirrel
12. hedgehog
13. snake
14. tortoise
15. gopher
16. cat
17. hare
18. otter
19. raccoon
20. badger
21. fox
22. dog
23. jackal
24. she-goat
25. he-goat
26. ewe
27. lynx
28. ram & mountain goat
29. wolf
30. boar
31. roe & donkey
32. human
33. mule
34. crocodile
35. deer
36. heifer
37. tiger
38. lion
39. cow & horse
40. stallion
41. mare
42. bull
43. ox
44. bear
45. camel
46. hippo
47. giraffe
48. elephant

Birds

1. sparrow
2. bullfinch
3. bat
4. swift
5. swallow
6. dove
7. jackdaw
8. cuckoo
9. magpie
10. quail
11. crow
12. partridge
13. hen
14. owl
15. rooster
16. eagle
17. crane & heron
18. stork

Utensils

1. needle
2. thimble
3. nail
4. awl
5. fork
6. spoon
7. scissors
8. knitting needle
9. knife
10. spindle
11. dagger
12. skimmer
13. hammer
14. rolling pin
15. scabbard
16. arrow & tongs
17. sickle
18. axe
19. saw
20. broom
21. pick axe
22. sledgehammer
23. saber
24. bow
25. gun
26. hoe
27. spade
28. staff
29. rake & scythe & pitchfork

Body parts

1. eyelash
2. tooth
3. fingernail
4. navel
5. eyebrow
6. eye
7. lip
8. gums
9. adam's apple
10. nose
11. finger
12. tongue
13. mouth & ear
14. chin
15. heel
16. elbow & cheek
17. jaw
18. fist
19. throat & forehead
20. palm of hand
21. hand & knee
22. neck
23. foot
24. shoulder
25. face
26. head
27. arm
28. leg
29. back (anatomical)

Note 4. Conceptual issues with size evaluation

In our study, we relate size-judgments made by the speakers of Russian and English to lexical items of languages spoken by Daghestanian highlanders. We took some measures to increase cross-cultural stability of size evaluation by excluding some of the concepts present in the original experimental (absolute size) and lexical (categorial size) data. All such exclusions were based on our personal judgments.

In the case of the Absolute Size Hypothesis, we excluded concepts which do not form part of the traditional lexicon of the languages we study. From the initial datasets (McRae et al. (2005), Binder et al. (2016), our own absolute size experiment), we excluded some culture-specific concepts which lack native equivalents in the languages of our study, e.g. such concepts as ‘anchor’. This concept was introduced in the highlanders’ languages not so long ago and is expressed by a recent Russian loanword. It is unlikely for the members of highlanders’ communities to encounter ship anchors in real life so size-based gender assignment rule, if present elsewhere, is most likely irrelevant for the word for ‘anchor’. At the same time, we did not exclude the concept ‘lighter’ from our absolute size datasets. It is also typically expressed with a recent loanword but is a well-familiar everyday object in the present-day highlanders’ villages and people are aware of its size characteristics. In principle, it may be a target of size-based assignment rule In the case of the Categorial Size Hypothesis, where the initial datasets were primarily based on the thesaurus in Kibrik and Kodzasov (1990), intended to reflect a Daghestanian’s real world experience. We excluded some concepts that we deemed less familiar to the Russian subjects who were mapping these concepts onto the scale of size. Examples of such concepts are Daghestanian fauna specimens, such as ‘hoopoe’ or some mustelids’ species.

Next, we are aware of the fact that size evaluation may be complex, and based on a set of different and partly independent parameters. The extension of an entity in one (stick), two (tabletop) or three (rock) dimensions may be judged differently in terms of size, probably making entities of different dimensions not intercomparable. Similarly, one could expect that flexibility (cf. flexible entities like rope or tablecloth or balloon as opposed to rigid entities such as a rod of the same length) may decrease the perception of size, while orientation in space (cf. upright entities lamp post or door or column as opposed to horizontal entities such as a log of the same length) may increase it; or at least that this may vary across languages.

We do not know which of the conceptual dimensions of size, and in which languages, are more prominent in gender assignment (cross-linguistically and in the languages in the study), if any size effects are to be observed at all. Our Russian subjects complained that ‘plate’ (two dimensions) is hard to compare with ‘awl’ (one dimension), so that, in the case of utensils, we profiled size in one dimension, which may be irrelevant for gender assignment in Archi. We currently see no way to overcome these problems, because scaling up to differentiate between dimensions of size as well as conceptually related categories such as weight requires far more massive experiments and lexical coverage.

Finally, it is far from always clear whether an experimentally elicited concept included in one of our English or Russian dataset finds a good match in East Caucasian lexicons. Across cultures, words like *bag* or *hat* may evoke very different concepts, and/or correspond to several concepts which lack a conventional hypernym. We explain how we dealt with such cases in the Note 5 of the supplement.

Even though we took some measures to address the problem of concept familiarity (see the previous section), the obvious flaw of our experimental design remains that we are trying to use experimental data obtained from speakers of one language (English or Russian in the case of absolute size; Russian in the case of categorial size) to test hypotheses about size effects on gender assignment in another set of languages. We cannot be sure that size judgments by speakers from such different cultural backgrounds – modern industrial societies on the one hand, pastoralists and sheepherders, on the other – are comparable. The main differential factor would probably be concept familiarity; highlanders probably have a more immediate experience with birds, animals and utensils than most Russian speakers.

For the categorial size experiment, we made a pilot run with 9 speakers of languages of Daghestan. These were not necessarily highlanders, not necessarily the speakers of the languages of our sample. We believe that the speakers of Daghestanian languages show a shared cultural and knowledge background to assume their size judgments are comparable. Daghestanian rankings showed moderate to strong similarity to the aggregated ranking based on the judgments of Russian speakers. We calculated Kendall’s rank correlation coefficients for both Daghestanian and non-Daghestanian subjects. For each individual, we computed the coefficient to indicate the strength of correlation between their decisions and the aggregated ranking. For the Russian subjects themselves, the correlation coefficient was calculated for the aggregated ranking excluding herself. The results are visualized below as Fig. 4:

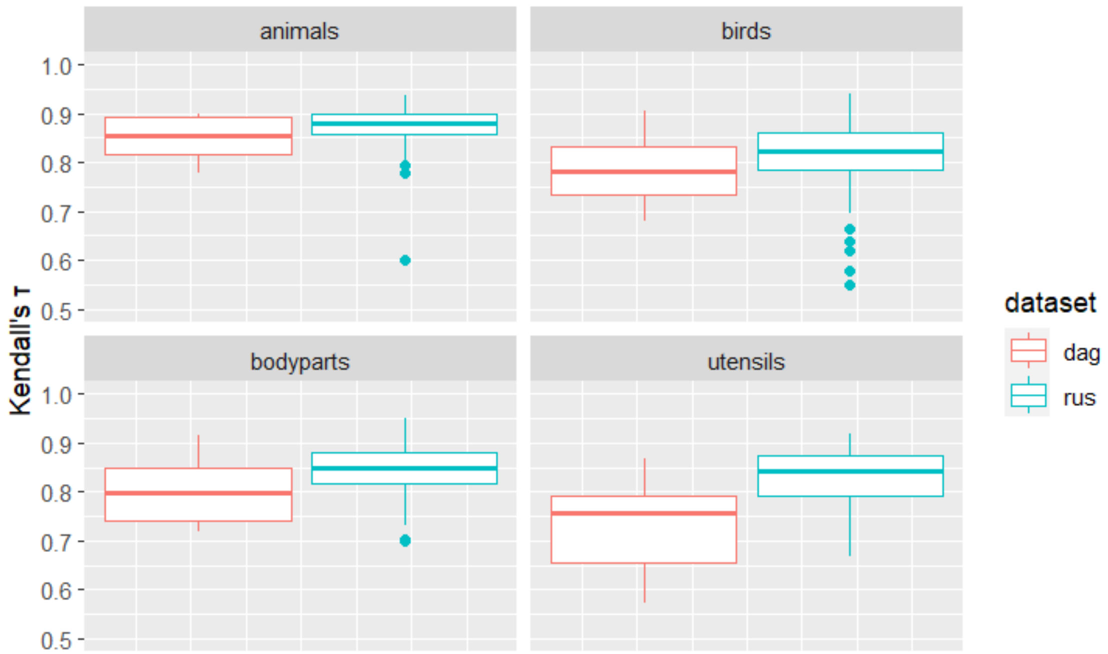

Figure 5. Comparison of the ranking carried out by Russian-speaking subjects (aggregated) with rankings by the speakers of East Caucasian languages

The boxplots show the range of variation of the correlation coefficients for Daghestanian (red) and Russian (blue) subjects. There is a slight tendency for the red boxplots to be displaced downwards; i.e. that an individual Daghestanian ranking may be slightly more divergent from the aggregated Russian ranking than an individual Russian ranking may be divergent from the aggregated Russian ranking, so differences in evaluation is an interesting outlook for further investigation (especially for body parts, where concept familiarity must be assumed to be the same). But on the whole the level of correlation seems to be high even in the red boxplots.

We did not try to corroborate the results of absolute size experiments by surveying native speakers of our target languages (or any East Caucasian languages) due to the nature of the data used in the absolute size experiment. Here, the size value of each concept is only meaningful when calculated from many individual rankings.

To sum up, the best target subjects for an independent replication of our experiments would be highlanders remaining in their natural environment (highest expected concept familiarity), but these are the most difficult to reach to run a large-scale online experiment. We could not run such a massive online survey of the speakers of our target languages. We believe that our pilot experiment for the categorial size corroborates relative cross-cultural stability of our rankings.

Note 5. Mapping concepts onto words

To investigate the relation between gender assignment in a lexicon of an East Caucasian language and experimental size judgments, we need to map concepts (represented by Russian or English words) from the latter onto words of the former. For the absolute size, we mostly used dictionary data, trying to stick to one dictionary per language: Alisultanov & Suleimanova (2019) for Rutul, Abdullaev (2018) for Lak, Chumakina et al. (2007) for Archi, Ibragimov & Nurmamedov (2010) for Tsakhur, Gummatov & Rind-Pawlowski (2020) for Kryz, Ganieva (2002) for Khinalug and Mejlanova (1984) for Budukh. For the data missing in the dictionaries, we consulted the speakers of the target languages (eliciting gender values either directly or in carrier phrases), especially in the case where a Russian loan is commonly used, i.e. Russian loan *sumka* which is a default word for 'bag' in Archi (other translations were highly culture specific and would not necessarily match size expectations for the referent of Russian *sumka*).

For the categorial size, we used the translations from Kibrik & Kodzasov (1989). With the exception of Rutul, we only did one lect per language (note that dialectal variation with East Caucasian languages may be very strong). In the case of Rutul, the following problem arises. Kibrik & Kodzasov (1989) describe the Luchek variety of Rutul. However, we could not use Luchek Rutul data to check both Categorial Size Hypothesis and Absolute Size Hypothesis since there is no dictionary of Luchek idiom from which we could potentially derive translations for concepts in our Absolute Size dataset. Thus, we decided to use the data from Mukhad Rutul - which is a standard variety documented by Alisultanov & Suleimanova (2019) - to be able to test for categorial and absolute size effects in the same variety of Rutul.

By definition, the input of semantic rules of gender assignment are meanings rather than forms of words. Our analysis must thus work with concepts rather than individual lexical items. Roughly speaking, we mapped gender values onto concepts via words for these concepts. This was important in the following case. A dictionary sometimes offers multiple translations for one concept. It is often unclear which word is a basic word for a concept, or which of the translation equivalents we found is its better match. In the absence of additional information, we had to consider these translations as full synonyms, even if they are not semantically identical.

In such cases, alternative translations may either belong to one or different gender values. When all translation equivalents of a concept belong to one gender value, the concept in the dataset is associated with this gender value, and was only counted once. However, if a concept has translation equivalents of both Gender 3 and Gender 4, we counted the concept twice, regardless of the actual number of equivalents of each gender. For example, Gummatov & Rind-Pawlowski (2020) offer both *xərə* (4) and *pəyə* (3) as Khinalug translation equivalents for a concept of 'pile', so that, for the purposes of statistical analysis of Khinalug data, this concept is counted once in Gender 3 and once in Gender 4. At the same time, of the two possible translations for 'pot' — *qəzənçə* and *qabləmə* — both belong to Gender 4, so this concept is only counted once.

These solutions were motivated by the following considerations. The assumption that semantic gender assignment works on concepts, not words, could potentially introduce a bias in our analysis by boosting the impact of this concept into the model. Assignment of two words for a small entity to Gender 4 or to Gender 3 may be seen as an unwarranted increase or decrease in detectability of the putative size effect, respectively, which makes our model more dependent on the impact from such concepts with one-to-many mapping. On the other hand, if several words for the same concept belong to different genders, we cannot know which of the assignments (if any) is semantically based, so we deemed it was safer to count the concept as ambiguous in terms of gender value (thus, it is possible that only one translational equivalent holds the influence of size-based gender assignment rules, and gender of another one is due to other factors). This solution requires the size effects in gender assignment to be stronger to be detectable, making the hypothesis more challenging to prove; the resulting correlation, if established, is more robust. Note that in most cases matching one concept onto multiple lexical items could be resolved through a discussion with a native consultant; but in practice that was only possible in the case of Archi.

On the other hand, if several concepts from our concept list were colexified in a target language, we mapped the gender of the corresponding lexical item on all such concepts. As a result, one gender value was associated with different sizes. For example, in Rutul, *mas* (4) was obtained as a translation equivalent of two concepts — 'wall' and 'fence', and Gender 4 was assigned to both concepts (Alisultanov & Suleimanova 2019: 260). In one case, however, we have merged what was initially considered to be different concepts into one. Since Russian lexically differentiates between 'heron', 'stork' and 'crane' and all three seem to be recognizable to Russian speakers, all three concepts were included into our birds dataset for the Categorial Experiment. However, later, these three were merged into one concept of 'long-legged bird' for two reasons. First, all three were grouped together on the scale of size, being the largest birds on the dataset. Second, the thesaurus in Kibrik and Kodzasov (1990) has only one concept for a long-legged, long-beaked bird. In their data for the languages of our study, there is only one word for such a bird, without further differentiation. Treating this as a systematic colexification of three distinct concepts with the same gender value would distort the statistical effect of what should rather be considered as only one concept.

Lastly, quite a few concepts are expressed in the languages in our study as compounds, including idiomatic multiword expressions (morphological compounds are not very typical of East Caucasian). We excluded from the final dataset for a language concepts that, in this language, were mapped onto a compound. To take one example, consider *q'alaq' uˤrbət'i* 'tortoise' (3). The second part of the compound *uˤrbət'i* means 'frog' (3), while the first part of the compound, not used on its own in Archi, is probably a loan from Lak *q'alaq'* 'lid, cover'). When testing for size effects, we cannot be sure whose size we are dealing with: the size of a Daghestanian tortoise (spur-thighed, aka Greek tortoise, whose length of the shell is on average between 12 and 15 cm) or that of a Daghestanian frog (probably Rana macrocnemis, whose length is less than 10 cm.) In our data, we did not find a single example where a compound did not have the same gender value as its head.

The process of mapping concepts onto words is summarized in Figure 5:

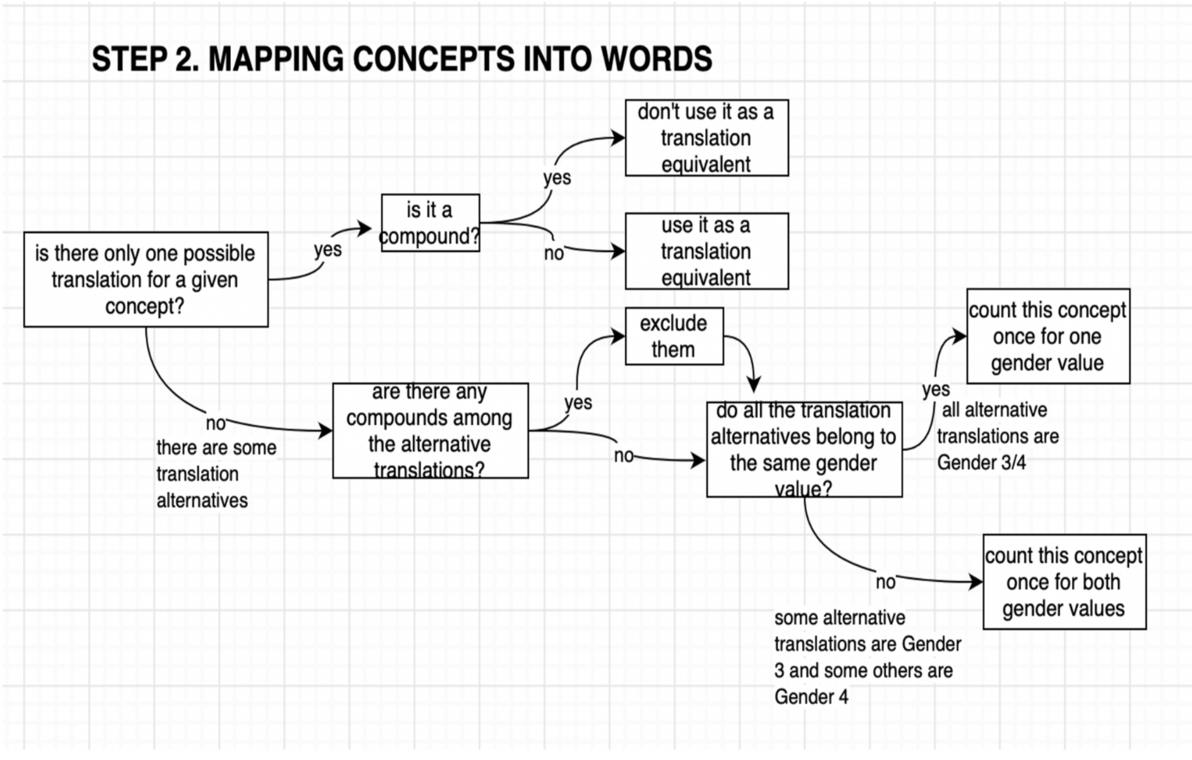

Figure 6. Mapping concepts into words

Note 6. Tsakhur: data on insects and animals

In terms of absolute size, Tsakhur is similar to other languages; the experimental data for the Absolute Size Hypothesis, when mapped onto Tsakhur lexical data, does not indicate presence of the absolute size effect. (It should however be noted that Tsakhur comes close to the significance level in testing the Hypothesis for big against McRae et al. 2005 dataset.) In the article, we have suggested that this may be due to the nature of the datasets we are using. The data related to categorial size effect, however, are worth a special discussion.

To start with the most straightforward case, Tsakhur only has one concept in the category of birds in Gender 4, and is thus similar to the other languages of the sample but not to Archi. This concept is 'bat' (which we put into this category in an attempt to approach the "naive" model of conceptualization, placing together all vertebrates that fly), which is indeed low on the scale of size and is mapped onto two lexical items in Tsakhur (*jarasa* and *čirχij*). Our approach to mapping requires that such datapoints are counted as one (see above the discussion in Note 5), which accounts for the absence of statistical significance, the only case of testing the Categorial Size Experiment where Mann-Whitney test could be applied (i.e., at least one concept is expressed by a word of Gender 4) but was not statistically significant (see Table 6). If the two lexical items for 'bat' are counted separately, p-value drops to within the selected level of significance (0.03).

In the discussion Kibrik and Kodzasov's (1990) data on insects in Section 6.3 on exceptional assignment, Archi groups with the other languages, while Tsakhur becomes the only language of the sample which has a considerable number of insects in Gender 4: 6 out of 19, including *zele* 'leech', *oˤlɨmčak* 'spider', *gindarang* 'spider', *kaba* 'butterfly', *c'itʲ* 'grasshopper', *šin* 'bedbug'. Here, however, the high number of Gender 4 nouns may have a more straightforward explanation. Our source gives two nouns for 'spider' and two separate nouns for 'butterfly' and 'grasshopper', which seem to be typical concepts for insects to go into Gender 4 (see Table 7 in the main article). In other languages, there is only one noun for the concept 'spider', and in Kryz, 'butterfly' and 'grasshopper' are reported as colexified. Note that the overall number of insects in Tsakhur is also higher than in any other language of the sample. We tentatively conclude that, in the category of insects, Tsakhur is not as different from the rest of the languages as it appears. Also note that for this category we do not have an experimentally established scale of size, so we cannot test whether Tsakhur Gender 4 insects are indeed deemed smaller that Gender 3 insects (to the authors' intuition, they are not), and that all languages including Tsakhur seem to have a pool of concepts for insects that tend to go to Gender 4. Cf. Table X, repeated here from the main article:

language	number of gender 4 insects (out of total)	insects
Archi	2/13	*žinžru* ‘leech’, *nat’* ‘nit’
Lak	2/15	*χ:*^{wʕ}*arcu* ‘spider’, *t’ɨmit’aj* ‘butterfly’
Rutul	3/17	*šamkal* ‘spider’, *t’ebebel* ‘butterfly’, *c’ɨc’* ‘grasshopper’
Tsakhur	6/19	*zele* ‘leech’, *o*^{ʕ}*lɨmčak* ‘spider’, *gindarang* ‘spider’, *kaba* ‘butterfly’, *c’it*^{j} ‘grasshopper’, *šin* ‘bedbug’
Budukh	0/16	-
Kryz	2/17	*гон* ‘butterfly’, *c’ɨlc’äng* ‘butterfly/grasshopper’
Khinalug	1/11	*nimc’* ‘louse’

Table 5 (Table 7 in the main article). Insects of gender 4 according to Kibrik & Kodzasov (1990)

The data on animals, however, place Tsakhur separately from all other languages in the sample. On the one hand, like Archi and unlike most languages of the sample, Tsakhur assigns a considerable number of names for animals in Gender 4 - 9 out of 42. Obviously, this is far from being the case of exceptional assignment. On the other hand, Mann-Whitney test delivers an insecure result (p-value is 0,086), well above the selected significance level. And indeed, it is far from obvious that the distribution of animals between two gender values correlates the scale of size used in our tests. Consider the following Table 6.

elephant	*fil*	3	crocodile	*timsah*	4	fox	*sɨ*^{ʕ}*wa*	3
giraffe	*žiraf*	3	mule	*ɢatir*	3	badger	*pɨrsiq’*	4
hippo	*begemot*	3	roe	*ǯiwur*	3	racoon	*enot*	4
camel	*dewa*	3	donkey	*a*^{ʕ}*male*	3	hare	*ɢ*^{ʕ}*ɨje*	3
bear	*s*^{j}*o*	3	boar	*wok*	3	cat	*bisːiː*	3

ox	*jac*	3	wolf	*ǯanawar*	3	tortoise	*baʁa*	3
bull	*boʁa*	3	capra caucasica	*k'on*	3	snake	*χoče*	3
mare	*mad^jjan*	3	ram	*ʁarg*	3	hedgehog	*kirpik^j*	4
stallion	*ajʁɨr*	3	lynx	*vašaq'*	4	squirrel	*mišawul*	4
cow	*zer*	3	ewe	*woq'ˤa*	3	mole	*krot*	3
horse	*balkan*	3	he-goat	*q'ɨna*	3	mouse	*q'ˤow*	3
lion	*aslan*	4	she-goat	*c'eʔ*	3	lizard	*baˤnek*	4
tiger	*pelhang*	4	jackal	*čaq'al*	3	frog	*q'ulbaʁa*	3
deer	*maral*	3	dog	*χ^we:*	3			

Table 6. Distribution of Tsakhur words for animals across the scale of size

Table 6 shows that there are animals identified as big in our categorial size experiment that belong to Gender 4, such as *timsah* 'crocodile', *pelhang* 'tiger', *aslan* 'lion', *vašaq'* 'lynx' (along with small animals such as *baˤnek* 'lizard', *kirpik^j* 'hedgehog', *s^juric'ma* 'squirrel (or another small rodent)', as well as relatively big animals such as *pɨrsiq'* 'badger'). One suspicion is that big Gender 4 animals may be loanwords. This hypothesis does not work at least for recent borrowing (or nonce borrowings) from Russian. We elicited names of elements missing in dictionaries by providing Russian stimuli where these nouns would control agreement and suggesting that a Russian word can be used if no native word comes to mind. Indeed, for hippopotamus, giraffe and mole our consultant used Russian words, all of which controlled Gender 3 agreement. Moreover, some of the early and more well-established loanwords from Iranian (*fil* 'elephant', *dewa* 'camel', *ǯanawar* 'wolf', *maral* 'deer' from Farsi) and Turkic (*dewa* 'camel') control Gender 3 agreement.

We have no explanation for the distribution of the names of animals between Gender 3 and Gender 4 and welcome other attempts at solving this. If size related rules are involved at all, they strongly interfere with other factors of gender assignment.

Note 7. MW test results

1. Absolute size experiment

source	p-value	language
macabre.big	0,44923858	Archi
macabre.small	0,81229138	Archi
binder.et.al.big	0,64100362	Archi
binder.et.al.small	0,85444322	Archi
mcrae.et.al.big	0,02607672	Archi
mcrae.et.al.small	0,70881113	Archi
macabre.big	0,58348021	Budukh
macabre.small	0,55713238	Budukh
binder.et.al.big	0,32216993	Budukh
binder.et.al.small	0,8216388	Budukh
mcrae.et.al.big	0,40252841	Budukh
mcrae.et.al.small	0,69909174	Budukh
macabre.big	0,10342271	Khinalug
macabre.small	0,37355898	Khinalug
binder.et.al.big	0,3306751	Khinalug
binder.et.al.small	0,89035871	Khinalug
mcrae.et.al.big	0,01407992	Khinalug
mcrae.et.al.small	0,84387507	Khinalug
macabre.big	0,68287454	Kryz
macabre.small	0,35132617	Kryz
binder.et.al.big	0,42707895	Kryz
binder.et.al.small	0,70955824	Kryz
mcrae.et.al.big	0,21609943	Kryz

mcrae.et.al.small	0,93798405	Kryz
macabre.big	0,00599944	Lak
macabre.small	0,91107435	Lak
binder.et.al.big	0,89621869	Lak
binder.et.al.small	0,32797456	Lak
mcrae.et.al.big	0,12202235	Lak
mcrae.et.al.small	0,73914583	Lak
macabre.big	0,09552331	Rutul
macabre.small	0,81716977	Rutul
binder.et.al.big	0,32737111	Rutul
binder.et.al.small	0,40063938	Rutul
mcrae.et.al.big	0,57192117	Rutul
mcrae.et.al.small	0,8805624	Rutul
macabre.big	0,28240495	Tsakhur
macabre.small	0,49325881	Tsakhur
binder.et.al.big	0,24823752	Tsakhur
binder.et.al.small	0,92632335	Tsakhur
mcrae.et.al.big	0,05294756	Tsakhur
mcrae.et.al.small	0,71559984	Tsakhur

2. Categorial size experiment

concept	p-value	language
animals	0,00880507	Archi
bodyparts	0,89998275	Archi
birds	0,00406601	Archi
utensils	0,95347278	Archi

animals		Budukh
bodyparts	0,5	Budukh
birds		Budukh
utensils	0,16329852	Budukh
animals		Khinalug
bodyparts	0,86603353	Khinalug
birds		Khinalug
utensils	0,95578998	Khinalug
animals		Kryz
bodyparts	0,28320853	Kryz
birds	0,01730528	Kryz
utensils	0,64223409	Kryz
animals	0,06387196	Lak
bodyparts	0,54964606	Lak
birds		Lak
utensils	0,25571724	Lak
animals		Rutul
bodyparts	0,49278669	Rutul
birds	0,02761713	Rutul
utensils	0,28299622	Rutul
animals	0,0672956	Tsakhur
bodyparts	0,35280223	Tsakhur
birds	0,10742347	Tsakhur
utensils	0,72696804	Tsakhur

References:

Alisultanov, A. S. and Sulejmanova, T. A. (2019). Rutul′sko-russkij slovar′ [Rutul-Russian dictionary]. ALEF, Makhachkala.

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

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*, 1849-1863.

Chumakina, M., Corbett, G., Brown, D., and Quilliam, H. (2007). A Dictionary of the Languages of the Archi villages.

Ganieva, F. A. (2002). Xinalugsko-russkij slovar′ [Khinalug-Russian dictionary]. DNC RAN, Makhachkala.

Gummatov, I. A. and Rind-Pawlowski, M. (2020). Qrızca-azərbaycanca-ingiliscə lüğət [Kryz-azerbaijani-english dictionary].

Ibragimov, G. X. and Nurmamedov, J. M. (2010). Caxursko-russkij slovar′ [Tsakhur-Russian dictionary]. Dagestanskij Gosudarstvennyj Pedagogičeskij Universitet, Makhachkala.

Kibrik, A. and Kodzasov, S. (1990). Sopostavitel’noe izučenie dagestanskix jazykov: Imja, fonetika [Comparative Study of Daghestanian languages: Noun, Phonetics]. Number v. 2. Izdatel’stvo Moskovskogo Universiteta [Moscow State University Publishing].

McRae, K., Cree, G. S., Seidenberg, M. S., and Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, 37(4):547–559.

Mejlanova, U. A. (1984). Buduxsko-russkij slovar′ [Budukh-Russian dictionary]. Nauka [Nauka], Moscow.