Supplementary Note:

Data Collection

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, 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. The steps are as follows:

1. Size annotation of concepts

1.1. Absolute Size

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. In these, for hundreds of different objects, their featural representation is experimentally collected.
We considered using a more recent and much larger 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 larger definitions/characteristics 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 datasets are described in Sections 4.1.1 to 4.1.3. Detailed information on how exactly the size 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 Mcrae et al. (2005, PAGES), Binder et al. (2016, PAGES). Figure 1 below 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 rank of a concept 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). Thus, for McRae et al. (2005) and for the Dataset Macabre, the minimal frequency value on the charts below 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 — something is associated with a particular semantic characteristic, i.e. with the smallness of size

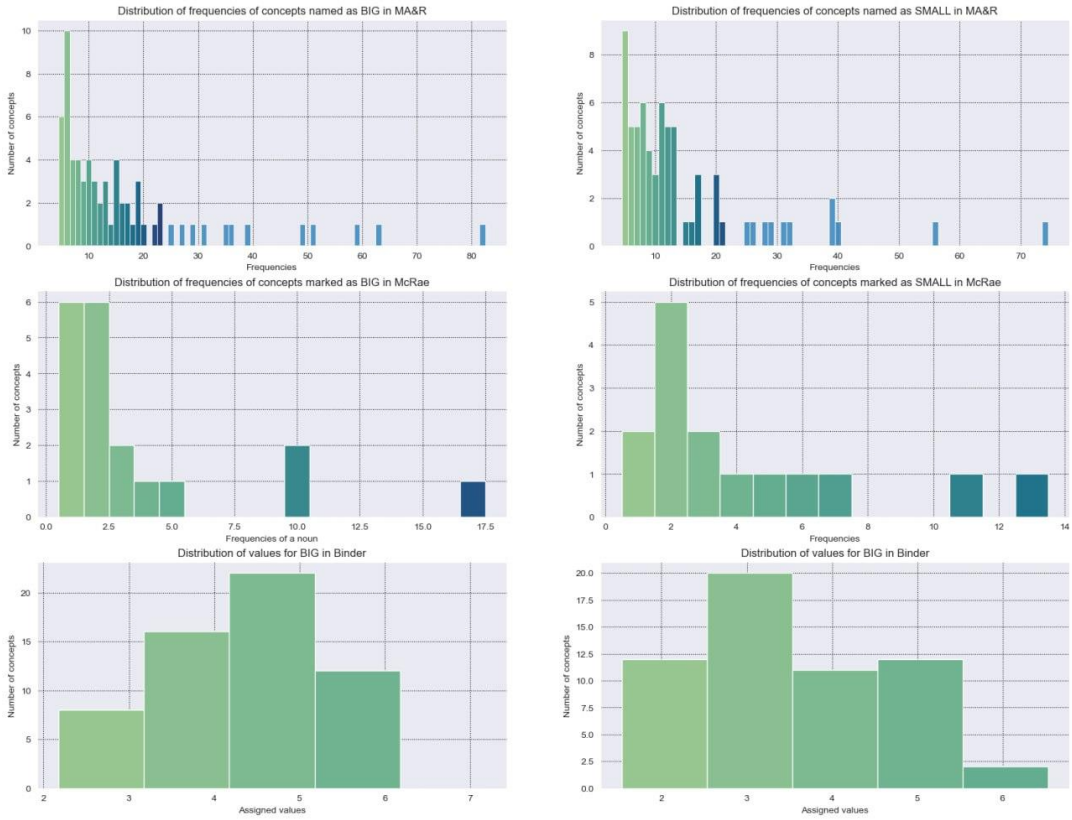

Figure 1. Distribution of different size values in three datasets.

The third dataset, Dataset Macabre, was compiled in an online experiment we ran ourselves. Fig. 2 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. Do not use singulative, diminutive or 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/.

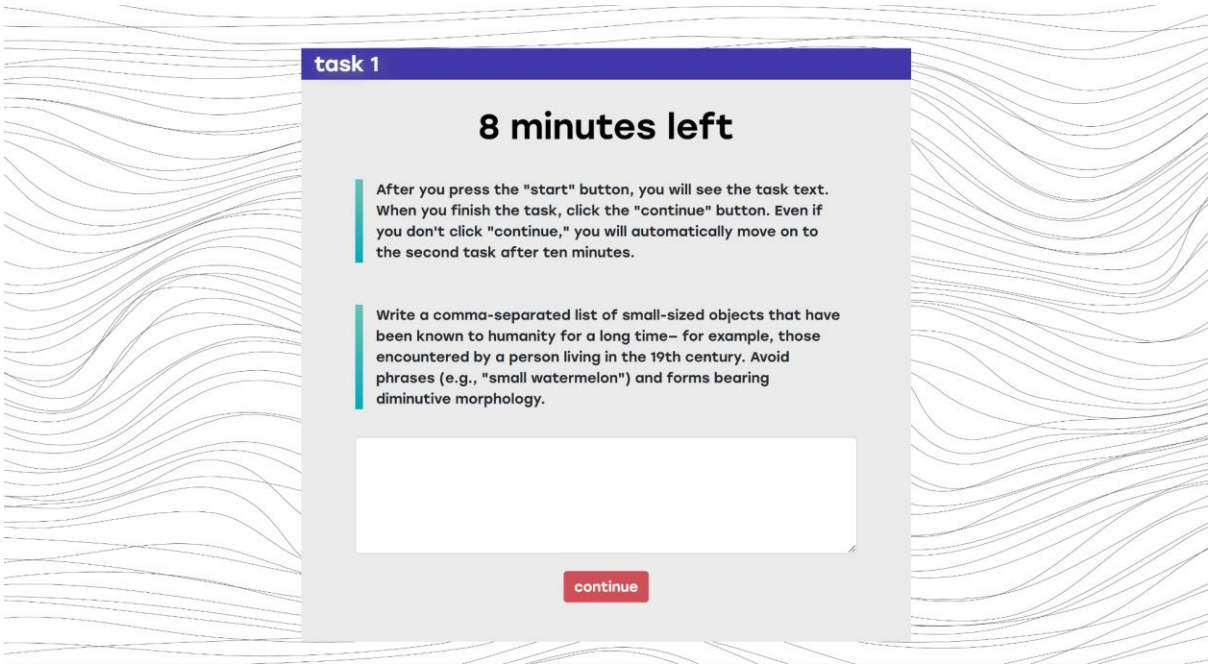

Figure 2. Interface of the absolute size experiment

1.1.1. 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.

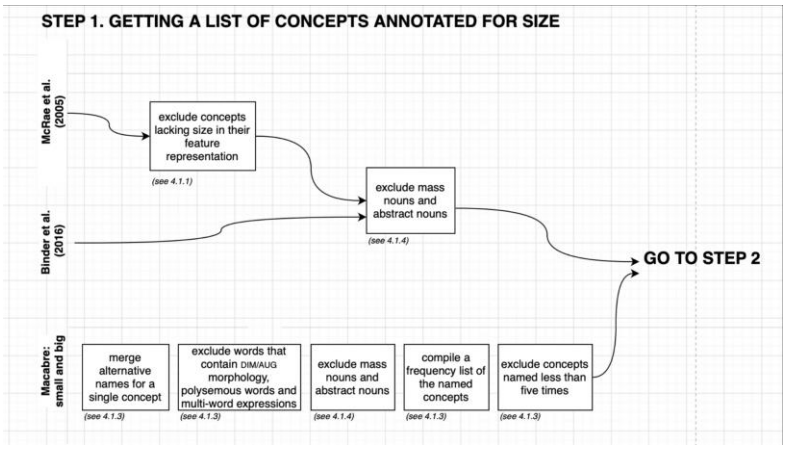

Figure 3. Getting a list of concepts annotated for size: the workflow

1.1.2. Comparison of the absolute size datasets

All concepts used in testing the Absolute Size Hypothesis are available on GitHub here. Section 4.1.4 of the paper briefly compares the datasets in terms of their content. Below, we provide a few more details. Figure 3 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.

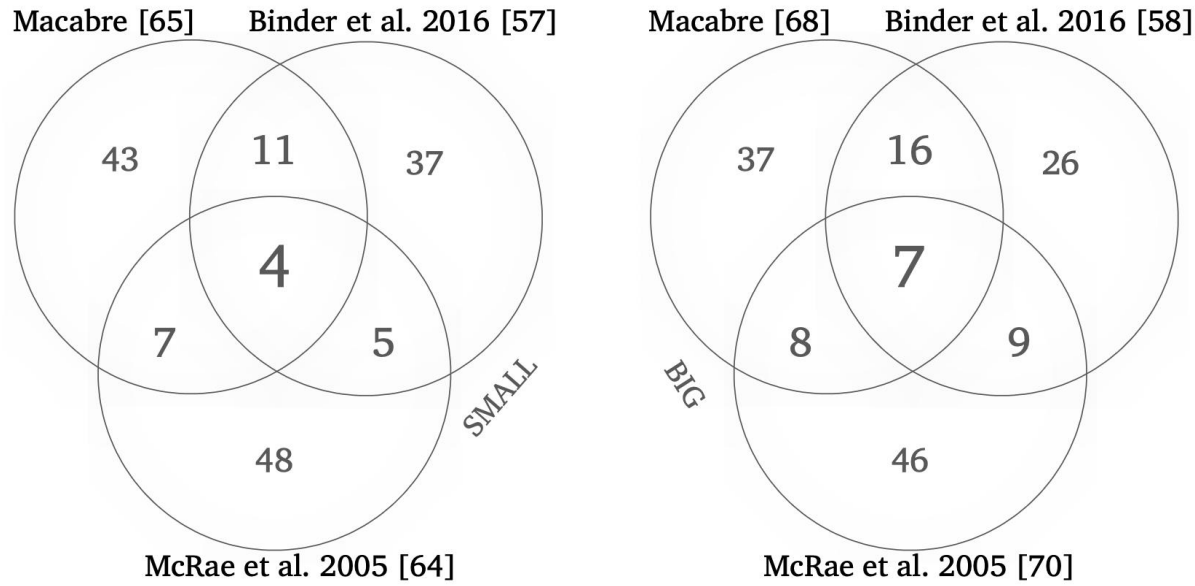

Figure 3. Overlap between three datasets

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 of such concepts
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 entitites present in more than one dataset

concepts	datasets	number of such concepts
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 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.

1.2. Categorial Size: issues with ranking

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 can be checked here (link). The raw results of the online-survey are available 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 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. After this, individual rankings became comparable. To generate an aggregated ranking, we computed the median of all rankings for each individual element. The overall aggregated rankings are the following (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

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

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)

An alternative approach to inter-respondent rank normalization is possible. This involves 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. To generate the aggregated ranking, we again computed medians for all elements. The overall alternative rankings are as follows.

Animals

1. ant
2. fly
3. spider & bee
4. butterfly
5. frog
6. mouse & lizard
7. bat
8. squirrel & hedgehog
9. snake & mole
10. gopher & tortoise

11. cat
12. badger & otter & raccoon & hare
13. dog
14. fox
15. jackal
16. she-goat
17. he-goat
18. mountain goat & ewe
19. wolf & boar & lynx
20. ram & donkey
21. roe
22. human
23. mule
24. crocodile
25. heifer
26. deer & tiger
27. stallion & mare & cow & lion
28. bull & ox & horse
29. camel & bear
30. hippo
31. giraffe
32. elephant

Birds

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

Utensils

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

23. scythe

Body parts

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

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 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.

1.3. Conceptual issues with size evaluation

One obvious flaw of our experimental design is 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 best target for such an experiment would be the highlanders remaining in their natural environment, but these are the most difficult to reach to run a large-scale online experiment.

One thing we could control is to exclude concepts which may not be familiar enough to the speakers of languages we study or to the Russian- speaking subjects who were assigning size values to the concepts. Such decisions were based exclusively on our own judgments. For the absolute size tests, where the initial datasets came from English subjects in McRae et al. (2005) and Binder et al. (2016) and from Russian subjects in our own experiment, we excluded some culture-specific concepts which lack native equivalents in the languages of our study, e.g. ship anchor or cigarette lighter. These concepts were introduced in the highlanders' communities not so long ago and are expressed by recent loanwords, mostly from Russian. We do not exclude all loanwords from the study — but we exclude 'anchor' and keep 'lighter' within our absolute size experiment. Lighter is a well-familiar everyday object in the present-day highlanders' villages and people are aware of its size characteristics. However, it is unlikely for them 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'. For the category size tests, where the initial datasets were primarily based on the Daghestan-oriented thesaurus in Kibrik and Kodzasov (1990), 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 bird or some mustelids' species only characteristic for this specific ecosystem.

Second, 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 and are most probably 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. We do not know which of the conceptual dimensions of size, and in which languages, are more prominent in gender assignment, if any size effects are to be observed. Thus, in 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 Section X of the supplement.

1.4. Daghestanian Rankings

We also made a pilot run of the categorial size experiment with 9 speakers of languages of Daghestan. These were not necessarily the speakers of the languages of our sample, but we believe that Daghestanian highlanders show a shared cultural and knowledge background to assume their size judgments are comparable. Daghestanian rankings showed moderate to strong correlation with the aggregated ranking based on the judgments of Russian speakers. More extensive discussion of Daghestanian categorial size rankings as compared to the aggregated ranking obtained from Russian speakers is present below in Section X of the supplement.

To substantiate the validity of the ranking derived from Russian speakers, we conducted a categorial experiment involving 9 speakers of languages of Daghestan. Subsequently, 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. The results were visualized through separate boxplots for Russian speakers and speakers of languages of Daghestan (Figure 4). These boxplots demonstrate partial overlaps, showing a moderate to strong correlation between Daghestanian rankings and the overall ranking across all datasets.

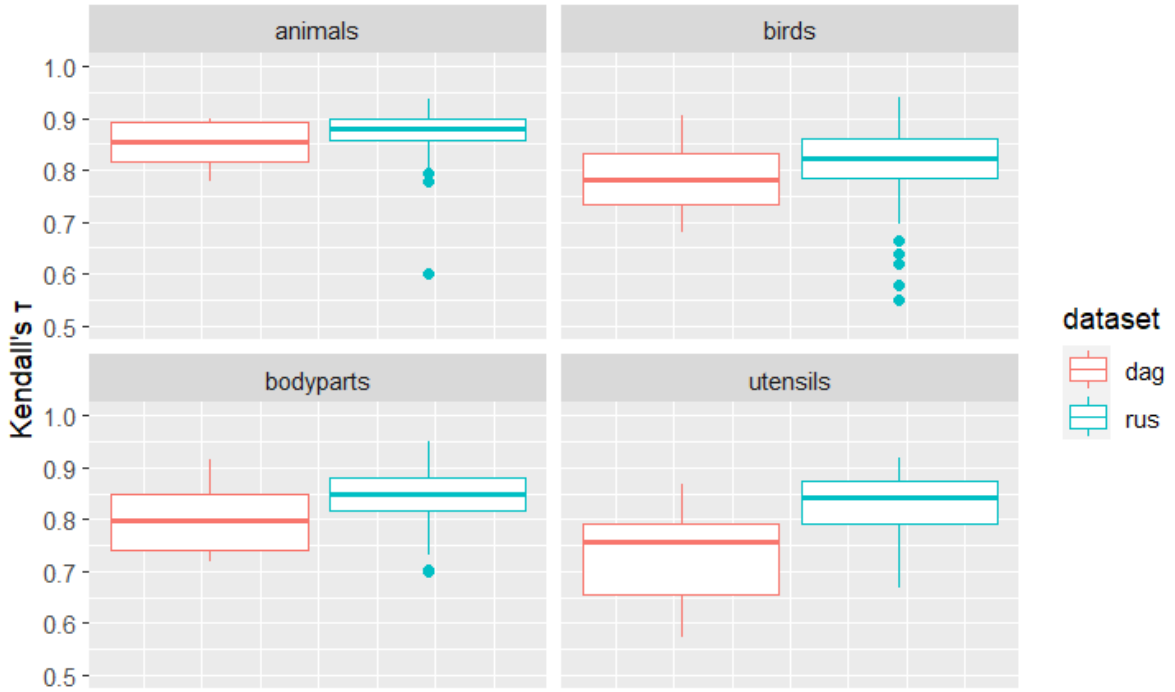

Figure 4. Comparison of the ranking carried out by Russian-speaking subjects (aggregated) with rankings by the speakers of East Caucasian languages

The least correlation is observed in utensils. It is likely to be attributed to the infrequent presence of most utensils in the day-to-day lives of the surveyed Russian speakers. In contrast, Daghestanian highlanders exhibit greater familiarity with these items. Despite this discrepancy, we maintain that our ranking remains satisfactory for the intended analysis.

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. We do not, however, have an opportunity to run an en masse online survey of the speakers of our target languages.

2. Mapping concepts into 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 Kibrik & Kodzasov (1989) as well as the speakers of the target languages (including 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 nowadays. For the categorial size, we used the translations from Kibrik & Kodzasov (1989)[2]. We only did one lect per language (note that dialectal variation with East Caucasian languages may be very strong).

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. 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. 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

[2] The only exception is Rutul. 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 Mukhad Rutul - which is a Standard variety documented by Alisultanov & Suleimanova (2019) - to test for categorial and absolute size effects together.

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 is semantically based, so we think 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 this was possible in the case of Archi much more often than for the other languages in the sample.

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 the languages of our study, there is always 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 as compounds in the languages in our study, including idiomatic multiword expressions (morphological compounds are not very typical of East Caucasian). Concepts that, in individual language, were mapped onto a compound, we excluded from the final dataset for this language. To take one example, investigating the impact of size on gender assignment for animals, we cannot be sure whose size we are considering in cases like *q'alaq' uʿrbət'i* 'tortoise' (3) - 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 (cf. *uˁrbət'i* 'frog' (3), where the first part of the compound, not used on its own in Archi, is probably a loan from Lak *q'alaq'* 'lid, cover'). In our data, we did not find a single example where a compound did not have the same gender value as its head. The translation proccess is summarized in Figure 5 below:

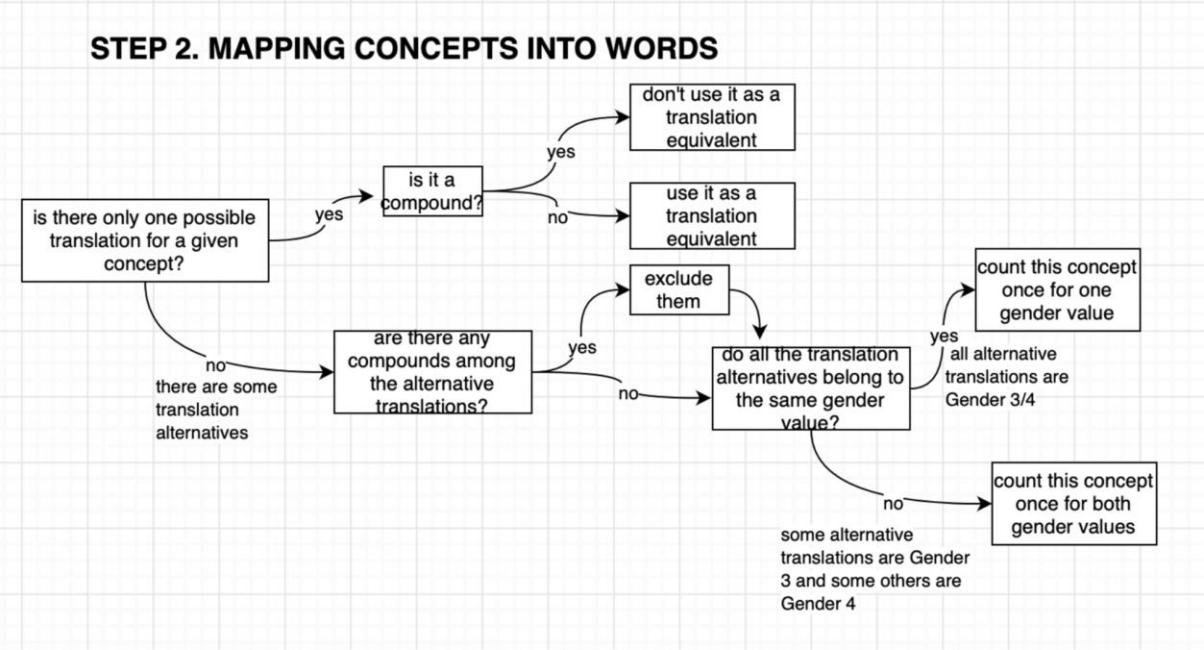

Figure 5. Mapping concepts into words