The Impact of Familiarity on Naming Variation: A Study on Object Naming in Mandarin Chinese

Yunke He[1], **Xixian Liao**[1], **Jialing Liang**[1] and **Gemma Boleda**[1,2]
[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra
[2]Catalan Institution for Research and Advanced Studies - ICREA
{xixian.liao,jialing.liang,gemma.boleda}@upf.edu
yunkehe66@gmail.com

Abstract

Different speakers often produce different names for the same object or entity (e.g., "woman" vs. "tourist" for a female tourist). The reasons behind variation in naming are not well understood. We create a Language and Vision dataset for Mandarin Chinese that provides an average of 20 names for 1319 naturalistic images, and investigate how familiarity with a given kind of object relates to the degree of naming variation it triggers across subjects. We propose that familiarity influences naming variation in two competing ways: increasing familiarity can either expand vocabulary, leading to higher variation, or promote convergence on conventional names, thereby reducing variation. We find evidence for both factors being at play. Our study illustrates how computational resources can be used to address research questions in Cognitive Science.

1 Introduction

When talking about objects in everyday experiences, people need to engage in the cognitive process of searching their lexicon to identify the most appropriate name to refer to them. This process involves intricate cognitive mechanisms that enable us to connect the properties of the object with the corresponding entries in our lexicon. Often, different individuals use different names to refer to the same object, reflecting the inherent variability in how we categorize and label our surroundings (Brown, 1958); for instance, the woman in Figure 1a can be called "woman", "tourist", or "person", among other choices. The reasons behind this variability are still not well understood.

Most previous research on naming has been done in Western languages (mostly English); and, in Cognitive Science, mostly with highly idealized stimuli, such as drawings of prototypical objects for a given category. Silberer et al. (2020b,a) introduced ManyNames, a dataset with realistic stimuli which provides an average of 31 English names for 25K objects in naturalistic images such as those in Figure 1. In this study, we present ManyNames ZH,[1] a new dataset for object naming that provides Mandarin Chinese names for a subset of the ManyNames data (1319 images, average 20 names per image). Figure 1 shows three example images with their corresponding names in ManyNames ZH.

We use this Language and Vision resource to address an open research question in Cognitive Science, namely, the role of object familiarity on naming variation. Familiarity is defined in psycholinguistic research as the level of prior exposure or knowledge that individuals have about specific stimuli, such as words and objects (Snodgrass and Vanderwart, 1980; Anaki and Bentin, 2009). We explore two seemingly opposite hypotheses, which respectively focus on two different aspects of naming variation: convergence on a conventional name, and size of the available vocabulary.

Hypothesis 1 (H1) posits that higher familiarity results in lower variation. This is based on the assumption that people tend to converge on a conventional name for familiar objects. Conversely, less familiar kinds of objects afford different conceptualizations, potentially increasing naming variation. For instance, most people are arguably more familiar with dogs than with bears, and indeed in Figure 1b Chinese subjects mostly converge on the majority name "狗" ("dog"), while they use a wider range of words to refer to the polar bear in Figure 1c. H1 has received support in some, but not all studies in Cognitive Science (see Section 2).

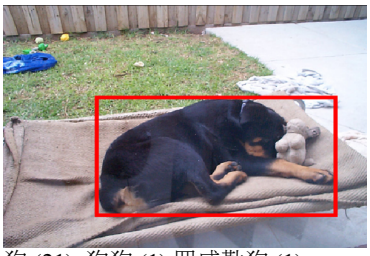
Hypothesis 2 (H2) instead suggests that higher familiarity is associated with increased naming variation. H2 is based on the idea that we need a larger vocabulary to refer to kinds of objects that we talk a lot about, to encode finer-grained distinctions in an efficient way (Gatewood, 1984). For instance, Silberer et al. (2020b) note that people elicit more

[1]Available at https://github.com/flyingpiggy1214/ManyNames_ZH

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 456–475
December 6–7, 2023.

女人 (12), 女士 (2), 人 (2), 大人 (1), 女 (1), 游客 (1)
woman (12), lady (2), person (2), adult (1), female (1), tourist (1)
Familiarity: 4.2 / H: 1.8 / N: 6

(a)



狗 (21), 狗狗 (1),罗威勒狗 (1)

dog (21), puppy (1), Rottweiler (1)

Familiarity: 4.1 / H: 0.5 / N: 3

(b)

北极熊 (8), 熊 (7), 动物 (2), 狗 (1), 海马 (1), 杂技 (1)
polar bear (8), bear (7), animal (2), dog (1), seahorse (1), acrobatics (1)
Familiarity: 2.5 / H: 2 / N: 6

(c)

Figure 1: Examples of images and their corresponding names in ManyNames ZH. Numbers in parentheses are counts across subjects. Familiarity is estimated by weighted average of lexical frequency (see section 4); H, or entropy, measures naming variation (see section 4); N is the number of distinct names.

variation than animals in ManyNames; according to H2, this would be due to the availability of a varied lexicon covering different dimensions that are relevant to categorize people, such as age ("child"), gender ("woman"), role ("tourist"), or profession ("lawyer"). A larger vocabulary means more naming choices, which then results in higher variation across subjects. The mirror argument applies to less familiar kinds of objects such as animals.

We find evidence for both hypotheses in our analysis of the ManyNames ZH data, and suggest how to reconcile the two.

2 Background

Object naming in Psycholinguistics and Cognitive Science. Naming an object involves the selection of a specific term to refer to it (Silberer et al., 2020a). In our daily life, it's common for objects to simultaneously fit into several categories; for instance, a given baby can belong to multiple overlapping categories like PERSON, FEMALE, BABY, and GIRL, among others. The names associated to these categories (e.g. "human", "person", etc.) are then all valid alternative names for this baby (Brown, 1958), resulting in variation. By far the most examined dimension of variation has been the taxonomic one, starting with seminal work by Rosch and colleagues (Rosch et al., 1976). This line of work divides categories into three levels: superordinate (e.g., ANIMAL), basic (e.g., DOG), and subordinate (e.g., ROTTWEILER). Rosch and subsequent work showed that, in general, people prefer names corresponding to the basic level, which is hypothesized to represent a good balance between the specificity and distinctiveness of the categories (Murphy and Brownell, 1985). However, another very prominent source of variation is so-called cross-classification (Ross and Murphy, 1999; Shafto et al., 2011), whereby objects belong to different categories that are not hierarchically organized but merely overlap (for instance, WOMAN and TOURIST).

In Cognitive Science, picture naming is the most widely used experimental paradigm for aspects related to naming (Snodgrass and Vanderwart, 1980; Brodeur et al., 2010; Liu et al., 2011; Alario and Ferrand, 1999; Tsaparina et al., 2011). Participants are presented with a visual stimulus and asked to produce the first name that comes to mind. The resulting datasets are called picture-naming norms, or naming norms for short. An important point for our purposes is the fact that, typically, due to the research goals of most of this research, the stimuli are prototypical pictures that represent categories, rather than the varied kinds of instances that one encounters in real life. Therefore, subjects reach a very high agreement in this task in terms of lexical choices (Rossion and Pourtois, 2004). This is also true for the few naming norms that exist for Mandarin Chinese (Liu et al., 2011; Weekes et al., 2007; Zhou and Chen, 2017). ManyNames (Silberer et al., 2020a,b) draws inspiration from this paradigm but uses real-world images that show objects in their natural contexts, which elicits much more variation.

Previous work has shown that properties related to lexical access (word frequency, age of acquisition) affect the production probability of names (Alario and Ferrand, 1999; Brodeur et al., 2010; Snodgrass and Vanderwart, 1980; Tsaparina et al.,

2011): All else being equal, more frequent words and words acquired earlier are preferred. Although less studied, research also shows that the properties of the pictured objects influence people's naming choices; objects that are less typical for the category denoted by the most produced name trigger higher variation (Snodgrass and Vanderwart, 1980; Gualdoni et al., 2022). People's naming choices are more varied for objects that are less typical for a frequent name. We focus on a different factor, namely familiarity (see below for more information).

Object naming in Computer Vision and Language & Vision. The task of Object Recognition in the realm of Computer Vision aims to identify and classify objects, assigning them a single ground-truth label from a pre-defined vocabulary (Everingham et al., 2015; Russakovsky et al., 2015; Kuznetsova et al., 2020). While this approach resembles picture naming, most of this research overlooks linguistic aspects related to natural language, in particular the fact that categories overlap and that different words can be used for a single category. The ManyNames dataset, from which we draw our images, was built a.o. as a response to this issue (Silberer et al., 2020b).

Several resources in Language & Vision (a field at the intersection between Computer Vision and Computational Linguistics) have collected referring expressions for real-world images. While existing resources like RefCOCO and RefCOCO+ (Yu et al., 2016), Flickr30K-Entities (Plummer et al., 2015), and VisualGenome (Krishna et al., 2017) can be a source naming data for objects in context, they lack sufficient data for a systematic assessment of the variability and stability of object naming. In contrast, ManyNames focuses on object names in isolation and elicits many more names for the same object from different subjects than any other resource to date.

Familiarity and naming behavior. In psycholinguistic research, traditionally familiarity has been assessed through rating tasks, where participants assign ratings on a scale to indicate the degree of familiarity they have with the stimuli (Snodgrass and Vanderwart, 1980; Sirois et al., 2006; Boukadi et al., 2016). Participants are instructed to consider objects encountered frequently in their daily lives as familiar, while categorizing rare or infrequently encountered objects as unfamiliar. In picture naming norms, familiarity, along with factors such as name agreement, lexical frequency, imageability, age of acquisition, and visual complexity, has been identified as a predictor of naming latencies[2] for both object and action pictures (Snodgrass and Vanderwart, 1980; Sirois et al., 2006; Liu et al., 2011). It has also been shown to affect lexical choice (Anaki and Bentin, 2009). For example, when presented with an object like Figure 2, individuals who describe it as "bread" or "burger" likely possess limited prior knowledge about different types of bread in the USA. On the other hand, if someone readily identifies the object as a "bagel", it suggests a higher level of familiarity.

Familiarity has also been related to vocabulary size for a given domain. In a study by Gatewood (1984), fifty-four American college students ranked their familiarity and knowledge about four semantic domains: musical instruments, fabrics, trees, and hand tools. They were asked to list all the categories of each domain they could think of in a free-recall task. The results showed that familiarity strongly predicts the size of salient vocabulary in each domain.

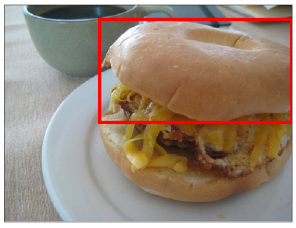


Figure 2: Image of a bagel.

The relationship between familiarity and naming variation, specifically, remains an open question, as results have varied across multiple studies. A large study of picture-naming norms (Krautz and Keuleers, 2022) found that naming agreement and accuracy were higher for those images that participants were familiar with. The same was found Tunisian Arabic data in Boukadi et al. (2016), and for Mandarin Chinese in (Liu et al., 2011; Zhou and Chen, 2017). However, a study of picture-naming norms for Canadian French by Sirois et al. (2006) revealed no relationship between naming agreement and object familiarity. Furthermore, note that familiarity has been shown to be culturally specific and may vary across different language communities (Boukadi et al., 2016). For instance, the Mex-

[2]The time it takes for a subject to start producing a name for a given stimulus.

ican dish guacamole may not be familiar within Chinese-speaking contexts.

In our study, we focus on the level of familiarity among Mandarin speakers regarding the objects sampled from the ManyNames dataset, and how this factor influences their naming variation. The stimuli thus are very different from the ones traditionally used in psycholinguistics, and can shed complementary light on the relationship between familiarity and naming variation. We also experiment with a corpus-derived measure of familiarity instead of using human ratings.

3 The ManyNames ZH dataset

3.1 Source dataset: ManyNames

Our ManyNames ZH dataset is based on the verified ManyNames dataset (ManyNames v2).[3] The original ManyNames dataset (Silberer et al., 2020a) provides 36 crowd-sourced annotations for 25K object instances obtained from VisualGenome (Krishna et al., 2017). The objects are categorized into seven domains: ANIMALS_PLANTS, BUILDINGS, CLOTHING, FOOD, HOME, PEOPLE, and VEHICLES. The annotations were obtained through an elicitation task conducted on Amazon Mechanical Turk (AMT), where participants were instructed to produce the first name that came to mind describing the object outlined by the red bounding box. To address the presence of noise in the data, a second version of ManyNames was created (Silberer et al., 2020b). Specifically, another round of annotation tasks was conducted on AMT to clean naming errors. Analysis revealed that most inadequacies correspond to referential issues (e.g., subjects responding "ball" for the image in Figure 1c; in Mandarin Chinese, no subject produced "ball", but instead they produced "acrobatics"). We used the English annotations to select a balanced sample of stimuli, as explained next.

3.2 Image sampling

ManyNames consists of 1319 images, sampled in 3 steps illustrated in Figure 3.

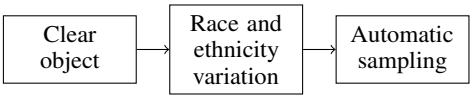


Figure 3: Image sampling procedure.

In Step 1, we filtered unclear images from ManyNames v2 to mitigate referential issues, keeping only images where at least 75% out of the subjects agree on the object being targeted.

In Step 2, we made an intervention in the PEOPLE domain to ensure variability in race and ethnicity within the selected images. The ManyNames dataset primarily represents Western culture, particularly American culture, so a simple random choice would produce mostly images of white people. We used Computer Vision models to determine the race of individuals in the images, in particular the OpenCV (Bradski, 2000) and Deepface (Serengil and Ozpinar, 2020) libraries. Given noise in the automatically identified images, two authors of the paper annotated the identified images of non-white people.[4] A third author resolved discrepancies (see details in Appendix B). Images identified as picturing Middle-Eastern, Latino Hispanic and Indian people resulted in low inter-annotator agreement. We therefore included only images of Black and Asian individuals. We further randomly sampled an equal number of images depicting white people, paired on the basis of sharing the same top name (name most frequently produced by the subjects in ManyNames; for instance, it was "woman" for the image in Figure 1a) and falling within the same variation band (see Step 3; also see Table 6 in Appendix B for statistics of the images). In total, we sampled 186 images in this step, with 93 non-white and 93 white individuals.

Most images in ManyNames have low variation; there is a prevalence of top names with mid-lexical frequency; and an imbalanced distribution across domains, with the majority of images belonging to the HOME domain (see Table 3 in Appendix A). Step 3 consisted in applying a sampling procedure to obtained a more balanced representation of naming variation, lexical frequency, and domains (details in Appendix B).[5]

[3] Available at https://github.com/amore-upf/manynames.

[4] The tools we use are trained with images in facial datasets (e.g., see Taigman et al. 2014). Generally, efforts are made to include clear and well-captured face images in these datasets. The human faces in our images are not always distinctly presented or complete, posing challenges for automatic identification using Computer Vision tools.

[5] We also noticed that there was an image with the topname "shoe" in the PEOPLE domain, and removed it.

3.3 Data collection

The collection of object names was obtained via crowdsourcing tasks on both Prolific[6] and AMT[7]. The 1319 images were randomly divided into 7 lists, with participants being assigned randomly to one of the 7 lists. On average, it took approximately 40 minutes for a participant to complete the entire experiment.[8] The experiment interface and the instructions for annotators are included in Appendix D.

We also collected demographic data about the participants (detailed information in Appendix C). They were 146 Mandarin Chinese native speakers (61 females, 82 males, 1 non-binary individual and 2 participants with unknown gender). They ranged in age from 18 to 50 years old, with 70% belonging to the 18-35 age group.

We experienced difficulties obtaining data from Chinese speakers from these platforms because they prevail in Europe and USA, but not in China. On Prolific, a small portion of participants answered the questions in Cantonese or even English. On AMT, when we filtered for Mandarin Chinese, very few participants could see the task, so we had to remove the filter, resulting in most responses being in English. In the end, we collected data from 370 participants on AMT but could keep only 17. This is an example of the difficulties involved in building datasets for languages other than English.

3.4 Post-processing

We post-processed the data to remove noise. First, we removed incorrect responses according to the criteria used in ManyNames. The four primary types of inadequate annotations are: referential ("named object not tightly in a bounding box"), visual recognition ("named object mistaken for something else it's not, as in bear-dog"), linguistic (such as "dear" for "deer") and others (Silberer et al., 2020b). We used Google Translate to convert the identified mistaken English names in ManyNames v2 to Mandarin and excluded matching responses from the Chinese data.

Second, we converted responses in Pinyin, the primary romanization system for Standard Mandarin Chinese, into corresponding Chinese characters. We also eliminated responses containing expressions for uncertainty e.g., "不知道" ("I don't know"), and removed punctuation and non-Mandarin words.

Third, we used spaCy POS (part-of-speech) tagging (Honnibal and Montani, 2017) to identify and remove adjectives in the responses, resulting in responses containing head words only, such as "狗"(dog) instead of "黑狗"(black dog) and "小狗"(little dog).

Lastly, in the CLOTHING domain, despite the post-processing in Step 1, we still noticed errors related to subjects referring to the wearers rather than the clothing item. This is a common issue; Silberer et al. (2020b) hypothesize that it is due to people being much more salient than clothes for humans. We created a list of names for the PEOPLE domain by collating all the responses, manually excluded those associated with clothing, and filtered responses in the CLOTHING domain according to the cleaned list. Note that despite this procedure some noise in the data remains, such as the name "杂技" ("acrobatics") for the image in Figure 1c.

3.5 Results

Table 1 presents descriptive statistics for the entire dataset as well as for each of the seven domains (see next section for how naming variation and familiarity were computed). There are clear differences in terms of naming variation across domains, with BUILDINGS, PEOPLE and CLOTHING having higher naming variation than FOOD, HOME, VEHICLES and especially ANIMALS_PLANTS. Instead, mean familiarity is similar across domains except for PEOPLE, with 3.9 compared to around 3.1 in other domains. The last column in Table 1 contains the comparable vocabulary size, obtained by randomly downsizing all domains to the smallest domain (sampling 136 images for all domains). Vocabulary size is largest in BUILDINGS and HOME; ANIMAL_PLANTS has the lowest vocabulary size.[9]

4 Analysis

Estimates for variation and familiarity. As standard in picture norms, naming variation for

[6] https://www.prolific.co/.

[7] https://www.mturk.com/.

[8] In addition to collecting free names, there was a second part of the experiment that collected names after seeing a classifier. This second set of data was for a different study.

[9] HOME is a heterogeneous domain, so it is expected to have a large vocabulary size. We instead have no explanation for the large vocabulary size in the BUILDINGS domain at present. Also note that, even though the domain is called ANIMAL_PLANTS, the vast majority of the images in that domain correspond to animals.

Domain	N±std	H±std	F±std	#Img	Voc. Size	Comp. Voc. Size
buildings	8.0±3.1	2.3±0.9	2.9±0.5	170	503	423
people	7.2±2.1	2.2±0.5	3.9±0.4	320	501	284
clothing	6.8±2.1	2.2±0.6	2.9±0.3	145	295	281
food	6.2±2.4	1.9±0.8	2.8±0.3	136	269	269
home	6.0±3.0	1.7±0.9	2.9±0.4	203	556	414
vehicles	5.4±2.7	1.6±0.8	3.3±0.5	191	334	259
animals_plants	4.1±2.2	1.2±0.7	3.1±0.5	154	212	192
all	6.4±2.8	1.9±0.8	3.2±0.6	1319	2670	2122

Table 1: Descriptive statistics for ManyNames ZH. Columns from left to right: domain, number N of distinct names per object (mean ± standard deviation); naming variation H (mean ± standard deviation)); familiarity F (mean ± standard deviation); total number of images (#Img); vocabulary size (total name types); comparable vocabulary size (total name types calculated by randomly subsampling 136 images from all domains).

objects was estimated in terms of the entropy H of the responses. Snodgrass and Vanderwart (1980) introduced this metric and defined as in Eq. 1, where k refers to the number of different names given to each object and p_i is the proportion of annotators giving each name.

$$H = \sum_{i=1}^{k} p_i \log_2 \left(\frac{1}{p_i} \right) \quad (1)$$

In this study, we use lexical frequency as a proxy for familiarity, based on the established positive relationship between familiarity and frequency (Boukadi et al., 2016; Tanaka-Ishii and Terada, 2011). We aim at modeling the familiarity of kinds of objects represented in the images. As mentioned in Section 2, in naming norms typically the objects are highly prototypical of a single named category. Instead, our stimuli are real-world images that are not always prototypical for a single salient category. We use the naming responses as proxies for the categories that a given stimulus belongs to, and define familiarity as the weighted average of lexical frequency, as defined in Eq. 2. Here N is the set of responses for a given stimulus, $f(n)$ is the corpus-based frequency of name n, and the weighting factor $p(n)$ the proportion of subjects that produced that name. Frequency (in logarithm of base 10) for names was extracted from SUBTLEX-CH, a subtitle corpus of Mandarin Chinese (Cai and Brysbaert, 2010). For names not found in the corpus, we assign the average frequency of the remaining names associated with that object to them.

$$F := \sum_{n \in N} f(n) \cdot p(n) \quad (2)$$

Regression model. We fitted a linear mixed-effects regression model with naming variation as the outcome variable and fixed effects for familiarity, domain, and their interactions. All predictors were centered so that the reference level for each predictor is the overall mean across all levels of that predictor. The inclusion of the domain as a fixed effect allowed for the examination of potential systematic variations in naming across different domains. The interaction between familiarity and domain was included to explore whether the relationship between naming variation and familiarity is domain-dependent. The lists assigned to participants were treated as random intercepts. All analyses were performed using Bayesian inference methods, using the brms-package (Bürkner, 2021) of R (version 4.3.0, R Core Team 2021).[10]

5 Results

Fixed effect estimates are shown in Table 2, where effects whose credible intervals (CI) do not cross 0 are boldfaced. The observed overall relationship between familiarity and naming variation aligns with H1: higher familiarity with a particular kind of object is associated with lower naming variation.

However, the model also suggests that variation is very different across domains. The domains, arranged in ascending order of naming variation, are as follows: ANIMALS_PLANTS, HOME, FOOD, VEHICLES, BUILDINGS, CLOTHING, and PEOPLE (see Figure 4 for a visualization of model predictions for domains). Recall from Table 1 that PEOPLE has the highest mean familiarity, and it also exhibits the highest model-predicted variation

[10] Model in brms syntax: H ∼ familiarity * domain + (1| list).

Variable	Estimate	Est. Error	95% CI
Intercept	1.81	0.06	[1.68, 1.94]
Familiarity	**-0.55**	0.05	[-0.65, -0.46]
Domain-animals_plants	**-0.72**	0.05	[-0.83, -0.61]
Domain-home	**-0.38**	0.06	[-0.49, -0.27]
Domain-food	**-0.24**	0.08	[-0.40, -0.07]
Domain-vehicles	**-0.12**	0.05	[-0.22, -0.03]
Domain-buildings	**0.27**	0.06	[0.15, 0.39]
Domain-clothing	**0.42**	0.07	[0.28, 0.56]
Familiarity: home	**-0.44**	0.11	[-0.65, -0.24]
Familiarity: food	-0.20	0.17	[-0.53, 0.13]
Familiarity: animals_plants	-0.19	0.11	[-0.40, 0.03]
Familiarity: buildings	0.01	0.11	[-0.21, 0.23]
Familiarity: vehicles	0.19	0.09	[-0.00, 0.36]
Familiarity: clothing	**0.55**	0.15	[0.26, 0.84]

Table 2: Estimates of fixed effects when predicting naming variation (H) as a function of familiarity, domain, and the interaction between familiarity and domain. The last column shows the credible interval. Effects with CIs that do not straddle 0 are boldfaced.

when holding other factors constant; and the converse for ANIMAL_PLANTS. This supports H2: for domains that we are highly familiar with, we develop a larger vocabulary, and more lexical choices result in higher variation.

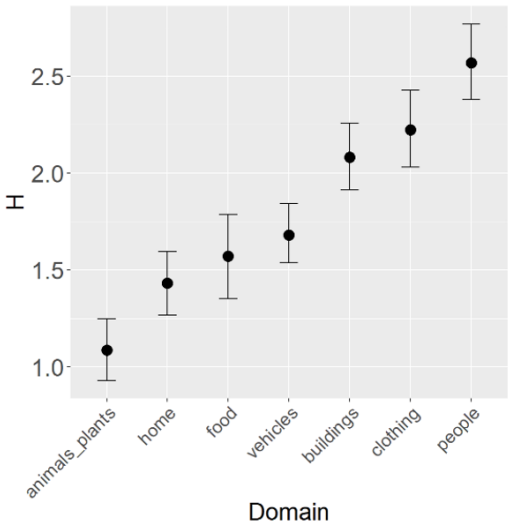


Figure 4: Predicted H of the domains covered in ManyNames ZH.

Furthermore, when examining the relationship between naming variation and familiarity across domains, we observe that CLOTHING is the only domain in which a higher familiarity of an object tends to increase, rather than decrease, naming variation.

6 Discussion

Our results suggest that, in general, higher familiarity predicts lower naming variation (Hypothesis 1) when Mandarin Chinese speakers name visually presented objects. This indicates that people tend to converge on a common name for kinds of objects they're more familiar with. For instance, in the ANIMALS_PLANTS domain, people exhibit relatively low naming variation when referring to dogs (see Figure 1b, where "dog" was produced by 21 out of 23 subjects). We hypothesize that this can be attributed to the prevalence of dogs as pets in our daily lives. Instead, we are less familiar with e.g. bears; in Figure 1c, people use "北极熊" ("polar bear") and "熊" ("bear") in almost equal proportion, and they also use the more general term "动物"("animal"). Note that some people do not correctly identify the kind of animal, naming it instead "狗" ("dog") or "海马" ("seahorse").[11]

However, an intriguing contradiction to this finding emerges when we consider the effect of different domains on naming variation. Although humans are arguably more familiar with people than with animals (conjecture supported by the data in Table 1), naming variation within the PEOPLE domain is actually much higher than that within the ANIMALS_PLANTS domain.[12] At the domain

[11] Silberer et al. (2020b) noted that subjects preferred the basic level term even if they risk being wrong (e.g. in cases where the gender of the person was not clear some subjects produced "man" or "woman" as opposed to "person").

[12] Silberer et al. (2020a) found the same for English.

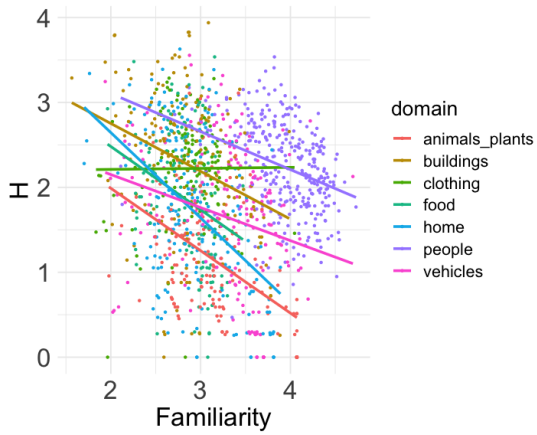


Figure 5: Effect by domain with a linear model.

Figure 6: Effect by domain using a GAM.

level, thus, naming variation actually *increases* with familiarity, in accordance with Hypothesis 2 and against Hypothesis 1. This is consistent with Gatewood (1984), which as discussed in Section 2 found salient vocabulary size to be positively correlated with familiarity in American English, for domains such as musical instruments. Chinese similarly seems to have a richer vocabulary for people as opposed to e.g. animals (see Table 1). This effect can be due to the fact that when we interact a lot with a given category of objects, like that of people, we need to develop a richer vocabulary to draw finer-grained distinctions within the category and facilitate communication. A larger vocabulary affords more opportunities for naming variation to arise.

Additionally, we also find evidence of the two factors being at play within the CLOTHING domain. While a linear regression model suggests that naming variation increases or plateaus in the CLOTHING domain (see Figure 5), fitting the data to a generalized additive model uncovers a clear convex curve (see Figure 6).[13] Manual inspection revealed that in the low-variation, low-familiarity area we have specific but unfamiliar objects like bowties; in the low-variation, high-familiarity area there are specific and familiar objects like t-shirts; and in the high-variation, mid-familiarity area there are types of clothes that are neither unfamiliar nor very familiar for Chinese speakers, like the jackets of masculine Western suits, which receive names such as “套装” and “西装” (“suit”), “衣服” (“clothes”), “外套” (“jacket”), or “西服” (“Western clothes”).

We thus find evidence for both hypotheses, which however play at different levels of granularity. At the level of a specific object, higher familiarity with that object’s category implies lower variation because people converge on the same label for the object. At the level of the domain or supra-category, instead, higher familiarity implies higher variation because of the richer vocabulary available for speakers.

7 Conclusion

In this paper, we have introduced ManyNames ZH, a new Language and Vision dataset designed for the task of Object Naming in Mandarin Chinese. The new dataset is the result of crowdsourcing names in Mandarin Chinese, based on the images from the English ManyNames dataset, with pre- and post-processing steps. ManyNames ZH consists of a carefully curated subset of 1319 images, each accompanied by an average 20 names provided by different human annotators. It allows the community to expand the empirical basis of findings on naming, by including a major language from a typologically different family than English. With the availability of ManyNames subsets in three languages, English, Catalan (Orfila et al., 2022), and Mandarin Chinese, researchers can also conduct cross-linguistic studies and comparative analyses on object naming.

With this new dataset, we have explored the relationship between object familiarity and the degree of naming variation. We observe two opposite factors at play. On the one hand, when familiarity with objects in a given supra-category or domain increases (such as with the PEOPLE domain), vo-

[13] The figure exhibits a smooth curve fitted to a scatter plot using geom_smooth() in ggplot2 (Wickham, 2016) with the method = “gam” argument and formula H ∼ s(familiarity, by = domain).

cabulary size correspondingly increases, too. This affords higher naming variation because it gives speakers more options to choose from. On the other hand, within a given category, more familiar sub-categories will afford conventionalization of the label used to talk about it, which elicits lower naming variation. This helps explain conflicting results found in Psycholinguistic studies on naming, which found the effect of domain on vocabulary size (Gatewood, 1984); a negative correlation between familiarity and variation variation (Krautz and Keuleers, 2022; Boukadi et al., 2016); and no relation between the two factors (Sirois et al., 2006), respectively.

Our analysis is based on a snapshot of Mandarin Chinese in which the vocabulary is frozen and we only observe the use. However, the patterns observed result from the dynamic evolution of vocabulary over time. Our results suggest that the need to frequently talk about a given kind of object triggers the development of a richer vocabulary that accounts for relevant distinctions within that broad class; and that higher communication about a specific kind of object triggers the convergence on a single label. Future work should test this hypothesis empirically.

Limitations

Our dataset still contains noise despite the post-processing efforts, particularly in the PEOPLE and CLOTHING domains. Challenges arise from referential errors, as well as the inclusion of non-noun words in the dataset. Additional steps, such as further semi-automatic or crowdsourcing-based filtering (as was done for the English ManyNames) could help address these issues.

Also, given the limited availability of native Mandarin Chinese speakers on the platforms we utilized, we were only able to gather an average of 20 annotations per image. In comparison, the English ManyNames dataset contains an average of 31 annotations per image. As mentioned above, this showcases the difficulties of building resources for non-Western languages.

It is also important to note that the images from the original ManyNames dataset primarily reflect the cultural background of the USA. We made an effort to balance racial representation in the PEOPLE domain, but we did not address cultural biases in other domains that are also heavily culture-dependent, in particular FOOD and CLOTHING, as we deemed it more difficult to do this with automatic means. Future work in Language and Vision needs to address cultural biases (Liu et al., 2021).

Finally, in our study, we used the weighted average of the lexical frequency of the responses as a measure of familiarity for objects. Alternatively, subjective ratings of familiarity by human participants can provide valuable insights and should be considered in future research. Also, there are individual differences in familiarity, and we provide a measure of overall expected familiarity within a culture, without taking into account these individual differences. We leave it to future work to investigate the relationship between familiarity and naming behavior at the individual level.

Ethics Statement

This paper complies with the ACL Ethics Policy. Quoting from the ACM Code of Ethics, we :(1) "contribute to society and to human well-being, acknowledging that all people are stakeholders in computing", by investigating how computational models can contribute to answer questions about how language works; (2) "avoid harm" by broadening the empirical basis of work on Language and Vision, introducing a new dataset for Mandarin Chinese; (3) are "honest and trustworthy" about our results and limitations; (4) "attempt to be fair and take action not to discriminate" by including considerations of race variability in our image sampling method (although future work should do more in including other sources of cultural variation); (5) "respect the work required to produce new ideas, inventions, creative works, and computing artifacts" by citing the related work that contributed to our work to the best of our knowledge; (6) "respect privacy" and (7) "honor confidentiality" by anonymizing the dataset prior to its public distribution. Like any work in AI and indeed in science and technology, of course, the results of our work can be used both for good and for bad.

Acknowledgements

This project has received funding from the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain; ref. PID2020-112602GBI00/MICIN/AEI/10.13039/501100011033). We also thank the financial support from the Catalan government (SGR 2021 00470) and the Department of Translation and Language Sciences at Universitat Pompeu Fabra.

References

F. Xavier Alario and Ludovic Ferrand. 1999. A set of 400 pictures standardized for french: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, 31:531–552.

David Anaki and Shlomo Bentin. 2009. Familiarity effects on categorization levels of faces and objects. *Cognition*, 111(1):144–149.

Mariem Boukadi, Cirine Zouaidi, and Maximiliano A Wilson. 2016. Norms for name agreement, familiarity, subjective frequency, and imageability for 348 object names in tunisian arabic. *Behavior Research Methods*, 48:585–599.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Mathieu Brodeur, Emmanuelle Dionne-Dostie, Tina Montreuil, and Martin Lepage. 2010. The bank of standardized stimuli (boss), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, 5:e10773.

Roger Brown. 1958. How shall a thing be called? *Psychological review*, 65(1):14.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Paul-Christian Bürkner. 2021. Bayesian item response modeling in r with brms and stan. *Journal of Statistical Software*, 100(5):1–54.

Qing Cai and Marc Brysbaert. 2010. Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6):e10729.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136.

John B Gatewood. 1984. Familiarity, vocabulary size, and recognition ability in four semantic domains. *American Ethnologist*, 11(3):507–527.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2022. Woman or tennis player? visual typicality and lexical frequency affect variation in object naming. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Agnieszka Ewa Krautz and Emmanuel Keuleers. 2022. Linguapix database: A megastudy of picture-naming norms. *Behavior Research Methods*, 54(2):941–954.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Youyi Liu, Meiling Hao, Ping Li, and Hua Shu. 2011. Timed picture naming norms for mandarin chinese. *PloS one*, 6(1):e16505.

Gregory L Murphy and Hiram H Brownell. 1985. Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of experimental psychology: Learning, memory, and cognition*, 11(1):70.

Mar Domínguez Orfila, Maite Melero Nogués, and Gemma Boleda. 2022. Cat manynames: A new dataset for object naming in catalan. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 31–36.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Brian Ross and Gregory L. Murphy. 1999. Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38:495–553.

Bruno Rossion and Gilles Pourtois. 2004. Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.

Patrick Shafto, Charles Kemp, Vikash Mansinghka, and Joshua B Tenenbaum. 2011. A probabilistic model of cross-categorization. *Cognition*, 120(1):1–25.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020a. Object naming in language and vision: A survey and a new dataset. In *Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the 12th Language Resources and Evaluation Conference; 2020 May 13-15; Marseilles, France. Stroudsburg (PA): ACL; 2020. p. 5792-801*. ACL (Association for Computational Linguistics).

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020b. Humans meet models on object naming: A new dataset and analysis. In *Scott D, Bel N, Zong C, editors. Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8-13; Barcelona, Spain. Stroudsburg (PA): ACL; 2020. p. 1893-905*. ACL (Association for Computational Linguistics).

Mélanie Sirois, Helgard Kremin, and Henri Cohen. 2006. Picture-naming norms for canadian french: Name agreement, familiarity, visual complexity, and age of acquisition. *Behavior Research Methods*, 38(2):300–306.

Joan G Snodgrass and Mary Vanderwart. 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

Kumiko Tanaka-Ishii and Hiroshi Terada. 2011. Word familiarity and frequency. *Studia Linguistica*, 65(1):96–116.

Diana Tsaparina, Patrick Bonin, and Alain Méot. 2011. Russian norms for name agreement, image agreement for the colorized version of the Snodgrass and Vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for modal object names. *Behavior Research Methods*, 43(4):1085–1099.

Brendan Stuart Weekes, Hua Shu, Meiling Hao, Youyi Liu, and Li Hai Tan. 2007. Predictors of timed picture naming in chinese. *Behavior Research Methods*, 39(2):335–342.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Dandan Zhou and Qi Chen. 2017. Color image norms in mandarin chinese. *Frontiers in Psychology*, 8:1880.

Appendices

A Image Sampling Statistics

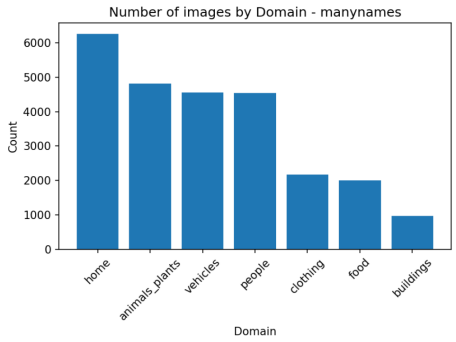
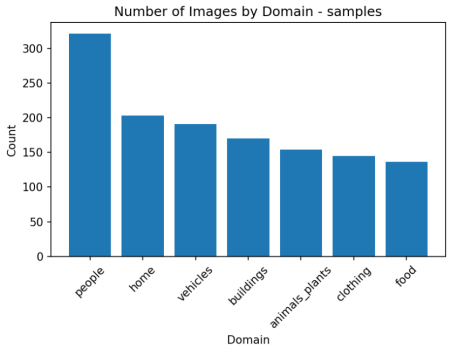
ManyNames v2	Sample
	

Table 3: Distribution of images across domains in ManyNames v2 and sample.

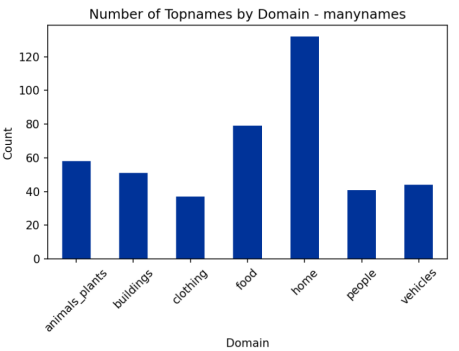
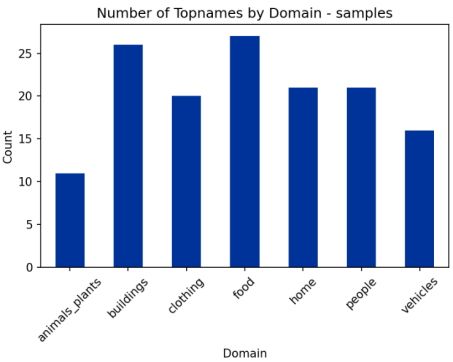
ManyNames v2	ManyNames ZH
	

Table 4: Distribution of topnames across domains in ManyNames v2 and ManyNames ZH.

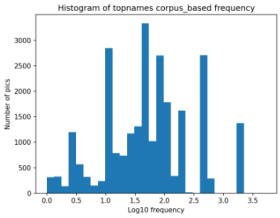
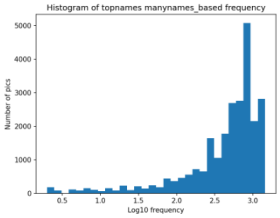
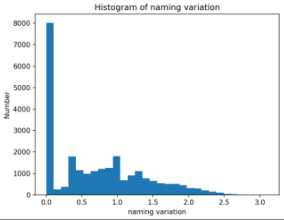
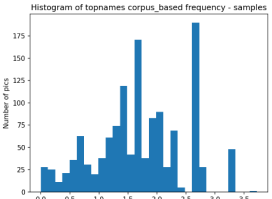
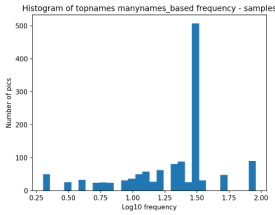
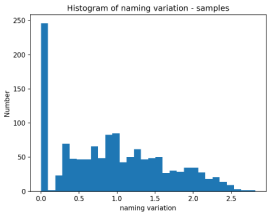
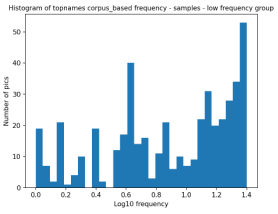
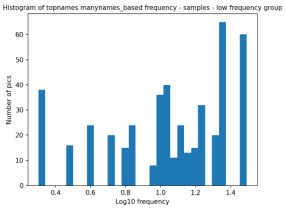
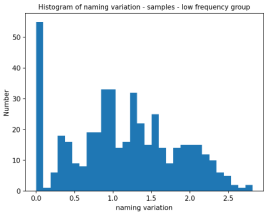
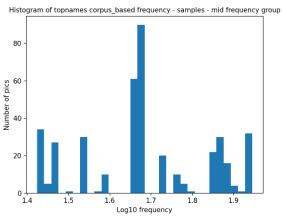
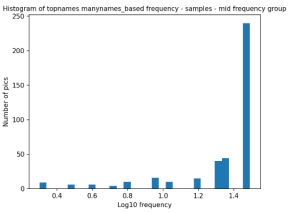
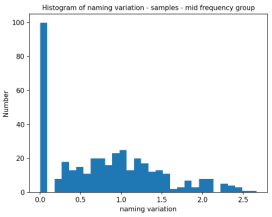
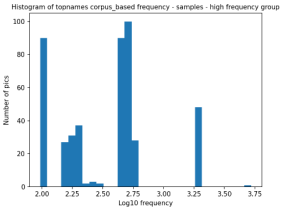
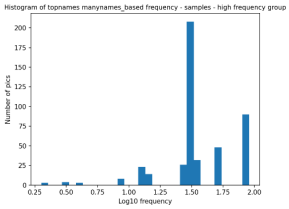
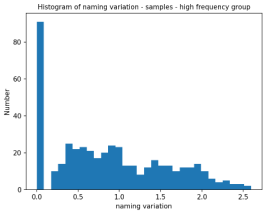
Dataset	Corpus-based frequency	ManyNames-based frequency	Naming variation
ManyNames v2			
Sample			
Sample-low frequency band			
Sample-mid frequency band			
Sample-high frequency band			

Table 5: Distribution of ManyNames, sampled images and each frequency band of sampled images in terms of topname frequency (corpus-based) in logarithm of base 10, topname frequency (ManyNames-based) in logarithm of base 10, and naming variation.

B Details on sampling

Table 6 shows the distribution of non-white images.

As for the automatic sampling, it consists of the following steps. First, we partitioned the images into three naming variation bands (low, mid, and high) using quantiles. Each band contained an equal proportion of the total images, resulting in approximately one-third of the images in each band. Likewise, we divided the topnames into three frequency bands (low, mid, and high) based on their corpus-based frequency in the logarithm of base 10 using quantiles. The frequency data were derived from SUBTLEX-US, a subtitle corpus of American English (Brysbaert and New, 2009). Each frequency band also contained approximately one-third of the topnames.

We initiated the image sampling from a specific domain (e.g., FOOD). Within the chosen domain, we focused on a particular frequency band (e.g., low frequency band). Next, we randomly selected a single topname (e.g., "cupcake") from the selected frequency band. For the chosen topname, we proceeded to sample 10 images from each of the low, mid, and high variation bands. If a variation band had fewer than 10 available images, we settled with all available ones and moved to the next variation band. We repeated this process of topname sampling until approximately 60 images were obtained for the selected frequency band. Following this, we repeated the sampling procedure for each frequency band within the selected domain, resulting in approximately 180 images obtained for each domain. This entire procedure was then replicated for the remaining six domains. Note that for the PEOPLE domain, we excluded previously sampled topnames from Step 2 to avoid duplication in this step (i.e., "woman", "man", "girl", "boy", "child" and "skier" in Table 6). We then sampled additional images until reaching 10 images or the maximum available per variation band. However, if the number of images for a specific topname already exceeded 10 in Step 2, we did not sample any additional images for that topname.

Race	Low	Mid	High
Asian	4 ("woman": 3, "man": 1)	38 ("woman": 27, "man": 9, "girl": 2)	39 ("woman": 9, "girl": 9, "boy": 9, "man": 6, "child": 5, "skier": 1)
Black	0	6 ("man": 4, "woman": 2)	6 ("boy": 2, "child": 2, "woman": 2)
Total	4	44	45

Table 6: Distribution of non-white images sorted by naming variation band; number out of parentheses is the number of images, and number in parentheses indicates the number of images with the corresponding top name.

C Demographics

Demographic questionnaire

中文物体命名：背景调查表

实验之前需要填写一份背景调查。相关信息严格保密的，不会以任何方式与您的姓名或身份相关联。请尽您所能回答问题。如果您对这份问卷有任何问题或疑虑，请在继续填写之前发送邮件到：*[email address]*

注意：标有星号（*）的问题是必答题。回答后才能进入下一步，谢谢您的合作！

1. 您的年龄？*
   （）18-25
   （）26-35
   （）36-45
   （）46及以上
2. 您的性别？*

   _____
3. 您的学历（包括在读）？*
   （）"高中及以下"
   （）"大专"
   （）"本科"
   （）"硕士研究生"
   （）"博士研究生及以上"
4. 普通话是你小时候学习的第一种语言吗？*
   （）是
   （）否
5. 在15岁之前，您是否都在中国居住？*
   （）是
   （）否
6. 您还会说其他语言吗？*
   （）是
   （）否

   如果是，请写出其他语言中最精通的语

言和对该语言的熟练程度（熟练程度供参考：入门、基础、中级、高级、母语）：
参考示例：英语，高级

_____

7. 在6岁之前，除了普通话之外，家里是否还有其他语言？*（包括方言）
（）是
（）否
如果是，家里说的是什么语言（或方言）：

_____

8. 您是否在非汉语国家学习或工作过？*
（）是
（）否
如果是，请说明居住时间最长的一个国家和大致居住的时间：
参考示例：西班牙，3年

Translation

Object naming in Mandarin Chinese: background questionnaire

A background survey needs to be completed prior to the experiment. The relevant information is strictly confidential and will not be associated with your name or identity in any way. Please answer the questions to the best of your ability. If you have any questions or concerns about this questionnaire, please send an email to *[email address]* before proceeding.

Note: Questions marked with an asterisk (*) are mandatory. Thank you for your cooperation!

1. How old are you? *(Required)

- 18-25
- 26-35
- 36-45
- 46 and above

2. What is your gender? *(Required)

_____

3. Please indicate your education level (including current status)* (Required)

- "High school or below".
- "Vocational college"
- "Bachelor's degree"
- "Master's degree"
- "Doctoral degree or above"

9. Was Mandarin Chinese the first language you learned as a child? *(Required)

- Yes
- No

10. Did you live in China until you were 15 years old? *(Required)

- Yes
- No

11. Do you speak any other languages? *(Required)

- Yes
- No

If yes, please write the most proficient of the other languages and the level of proficiency in that language (proficiency level for reference: Beginner, Basic, Intermediate, Advanced, Native): Reference Example: English, advanced

_____

12. Before the age of 6, were there any other languages spoken at home besides Mandarin (including dialects)? * (Required)

- Yes
- No

If yes, what language (or dialect) was spoken at home: _____

13. Have you ever studied or worked in a non-Chinese speaking country? *(Required)

- Yes
- No

If yes, please indicate the country where you have lived the longest and the approximate length of residence: Reference example: Spain, 3 years

_____

Variable	Category	Frequency	Percentage
Age	18-25	44	30.1%
	26-35	58	39.7%
	36-45	31	21.2%
	46-50	13	8.9%
Gender	Female	61	41.8%
	Male	82	56.2%
	Non-binary	1	0.7%
	Unknown	2	1.4%
Educational level	High school or below	3	2.1%
	Vocational college	8	5.5%
	Bachelor's degree	58	39.7%
	Master's degree	53	36.3%
	Doctoral degree or above	24	16.4%
Mandarin Chinese as first language learned?	Yes	139	95.2%
	No	7	4.8%
Live in China until 15 years old?	Yes	120	82.2%
	No	26	17.8%
Speak any other languages?	Yes	143	98.0%
	No	3	2.0%
Before the age of 6, were there any other languages spoken at home besides Mandarin (including dialects)?	Yes	70	48.0%
	No	76	52.0%
Have you ever studied or worked in a non-Chinese speaking country?	Yes	131	89.7%
	No	15	10.3%
n = 146			

Table 7: Descriptive statistics on the demographics of the participants in ManyNames ZH.

D Experiment Procedure

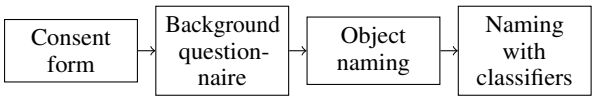


Figure 7: Experiment design

Our experiment consisted of four sessions: consent form, background questionnaire, object naming, and object naming with classifiers. The last one, adapted from the third session, served for another study.

Also, the initial pilot studies revealed that participants tended to use modifiers and numerical classifiers when describing objects. To address this, the instructions were modified to discourage the use of such linguistic elements. (see Appendix D for experiment interface and instructions for annotators).

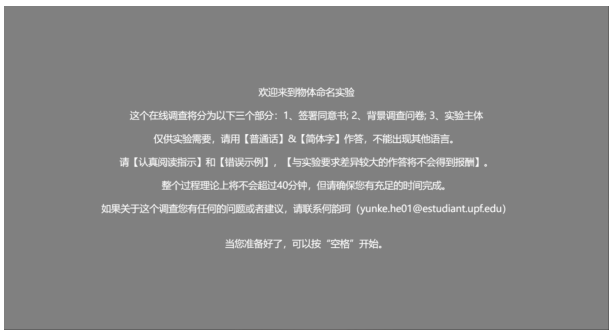


Figure 8: Introduction

Translation for Figure 8

Welcome to the object naming experiment.

This online survey is comprised of three parts: 1. Consent form; 2. Background questionnaire; 3. The main study.

Just for the purpose of the study, please answer all questions in Mandarin Chinese and Simplified Chinese; other languages are not allowed.

Please read the instructions carefully and the mistake examples carefully. No reward will be paid for answers that differ significantly from the experimental requirements.

Theoretically, the whole process will take no more than 40 minutes, but make sure you have enough time to finish this before you start.

If you have any doubts or questions about this study, please send an email to *[email address]*.

You can press [space] to start the experiment whenever you are ready.

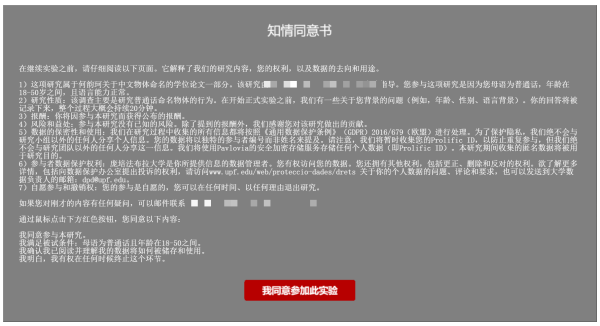


Figure 9: Informed Consent Form

Translation for Figure 9

Before you proceed with the experiment, please read carefully the following page. It explains our research, your rights, where the data goes, and what it is used for.

1. The experiment belongs to *[name]*'s study, supervised by *[name]*. You participate in this study because your native language is Mandarin Chinese, age is between 18-50 years old, and you have normal language ability.

2. Research description: This experiment mainly studies behavior for naming objects in Mandarin Chinese. Before the main experiment, we have some questions about your background (including age, gender, and language backgrounds). Your answer will be recorded, and the process will last approximately 40 minutes.

3. Reward: You will be paid with the published compensation.

4. Risks and benefits: Participation in the study entails no unknown risks. Besides the reward mentioned before, we appreciate your contribution to our study.

5. Privacy: All the information we collect during the course of the research will be processed in accordance with Data Protection Law. In order to safeguard your privacy, we will never share personal information with anyone outside the research team. Your data will be referred to by a unique participant number rather than by name. Please note that we will temporarily collect your Prolific ID to prevent repeated participation; however, we will never share this information with anyone outside the research team. The anonymized data collected

during this study will be used for research purposes.

6. Rights of participants: Pompeu Fabra University is the manager of your data. You have the rights to access your data, including correcting, deleting, and rejecting it. If you want to know more, please access www.upf.edu/web/proteccio-dades/drets. With respect to issues of personal data, you can also send an email to the responsible person of the university: dpd@upf.edu

7. Voluntary nature of participation: Your participation in this study is on a voluntary basis, and you may withdraw from the study at any time without having to justify why.

By clicking on the red button below, you agree to the following contents:

- I agree to participate in this study.
- I meet the criteria of participation: my native language is Mandarin Chinese, and my age is between 18-50.
- I confirm that I have read all the information above and understand how my data is going to be conserved and used.
- I understand that I have the right to terminate this study whenever I want.

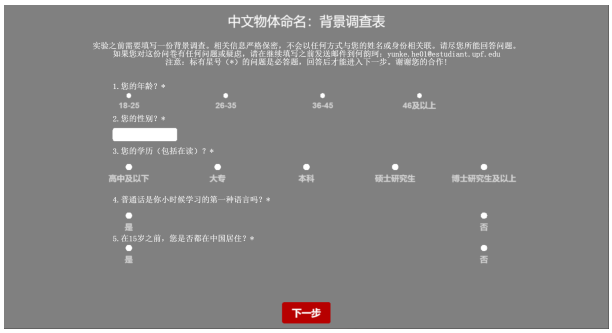


Figure 10: Background Survey(A)

Background survey is translated above in appendix C.

Translation for Figure 12

Welcome to our study! In the experiment, you will see about 250 images (200 for the first part and 50 for the second part), as shown in the figure. Your

Figure 11: Background Survey(B)

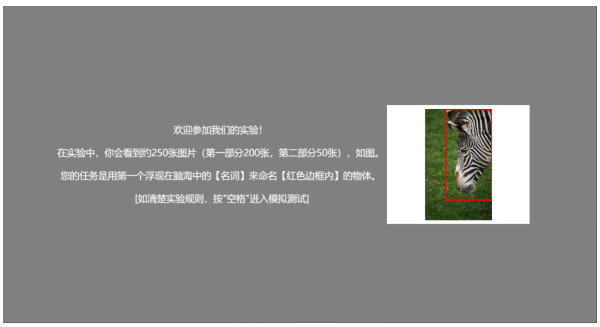


Figure 12: Part 1 Introduction

task is to name the object in the red bounding box with the first noun that comes to mind.

If you understand the rules, please press [space] to go to the next step.

Figure 13: Mistakes Exemplified in Part 1

Translation for Figure 13

Task: Please name the object in the red bounding box with the first noun that came to mind. Please read the instructions carefully and the mistake examples carefully. No reward will be paid for answers that differ significantly from the experimental requirements.

1. If multiple objects appear in the red bounding box, the object you should name is the most complete one in the bounding box.

2. Please try to avoid the mistakes exemplified (modifiers for color, status, and number) and fill in the input box as instructed on the right side.

Wrong answer: The upper part of the human body
Your answer:
Error cause: The red bounding box indicates the clothes, not the upper part of the human body
Right answer (just for reference): jacket, clothes...
Wrong answer: red car
Your answer:
Error cause: "red" refers to the color and has no relation to the object itself
Right answer (just for reference): car, taxi...
Wrong answer: the birthday girl
Your answer:
Error cause: "birthday" refers to the status of the girl and has no relation to the object itself
Right answer (just for reference): child, girl...
Wrong answer: a piece of cake
Your answer:
Error cause: "a piece of" describes the number and has no relation to the object itself
Right answer (just for reference): cake, cheese-cake

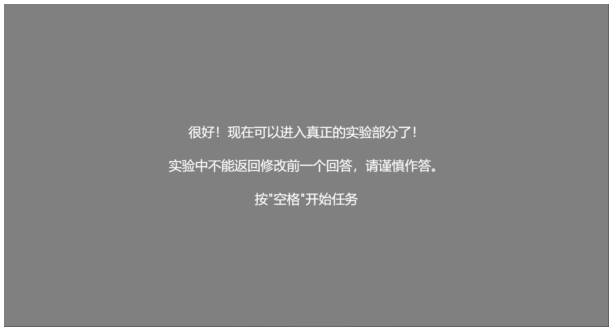


Figure 14: Notification for Starting Experiment

Translation for Figure 14

Great! Now you can go to the real experiment.

In the experiment you cannot go back to change the previous answer, please answer with caution.

Press [space] to enter the experiment.

Translation for Figure 15

Please name the object in the red bounding box with the first noun that came to mind and press [enter] to go to the next image.

Important: avoid modifiers for color, status and number; avoid usage of any verbs and adjectives.

Figure 15: Part 1 Object Naming Example

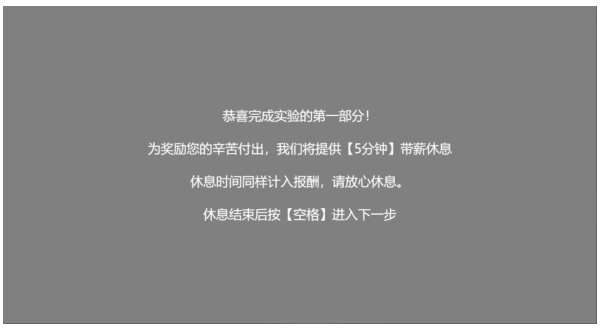


Figure 16: 5-minute-break between Part 1 and Part 2

Translation for Figure 16

Congratulations! You have finished the first part of the experiment!

To reward your hard work, we provide you with five-minute break with compensation included. Please take a rest.

After the break, you can press [enter] to go to the next step.

Figure 17: Part 2 Introduction

Translation for Figure 17

The second part of the experiment contains 48 images.

Your task is to name the object in the red bounding box with the first noun that came to mind, combing the classifier we give.

If you understand the rules, please press [space] to go to next step.

Figure 18: Mistakes Exemplified in Part 2

Translation for Figure 18

Task: please name the object in the red bounding box with the first noun that came to mind, combining the classifier we give.

1. If multiple objects appear in the red bounding box, the object you should name is the most complete single one in the bounding box.
2. Please try to avoid the mistakes exemplified (modifiers for color and status) and fill in the input box as instructed on the right side.

Wrong answer: one liang of [red car]
Your answer:
Error cause: the red indicates the color, has no relation to the object itself.
Right answer (just for reference): car, taxi...
Wrong answer: one piece of [sliced cake]
Your answer:
Error cause: sliced indicates the status, has no relation to the object itself.
Right answer (just for reference): cake, cheesecake...

Figure 19: Part 2 Object Naming with Classifier Example

Figure 20: End

Translation for Figure 19

please name the object in the red bounding box with the first noun that came to mind, combing the classifier we give, and press [enter] to go to the next image.

Translation for Figure 20

Thanks a lot for your participation!
Press [space] to exit.