CAT ManyNames: a New Dataset for Object Naming in Catalan

Mar Domínguez Orfila
Universitat Pompeu Fabra
Barcelona Supercomputing Center
mar.dominguez01@estudiant.upf.edu

Maite Melero Nogués
Barcelona Supercomputing Center
maite.melero@bsc.es

Gemma Boleda Torrent
Universitat Pompeu Fabra
gemma.boleda@upf.edu

Abstract

Object Naming is an important task within the field of Language and Vision that consists of generating a correct and appropriate name for an object given an image. The ManyNames dataset uses real-world human annotated images with multiple labels, instead of just one. In this work, we describe the adaptation of this dataset (originally in English) to Catalan, by (i) machine-translating the English labels and (ii) collecting human annotations for a subset of the original corpus and comparing both resources. Analyses reveal divergences in the lexical variation of the two sets showing potential problems of directly translated resources, particularly when there is no resource to a proper context, which in this case is conveyed by the image. The analysis also points to the impact of cultural factors in the naming task, which should be accounted for in future cross-lingual naming tasks.

1 Introduction

Most NLP resources are only available for a small percentage of languages (Joshi et al., 2020), being the rest of the languages spoken in the world left behind. This affects also Catalan, which can be considered a moderately under-resourced language (Armengol-Estapé et al., 2021). In the Language and Vision area, although significantly large datasets of annotated images have been created for a variety of tasks for English, to date no resources of this kind exist for Catalan. In this work we present CAT ManyNames[1], the Catalan version of the ManyNames dataset, which is the first available resource for the task of Object Naming in Catalan. The dataset has been translated from the English version and its test set has been human annotated to assess the quality of the translation. We also provide analyses of the sources of variation between the human annotated dataset and its translated counterpart.

2 Background

2.1 Object Naming: an Interdisciplinary Task

Naming an object accounts for picking out a nominal to refer to it (Silberer et al., 2020a), and is a linguistic phenomenon that can show lexical variation. On the one hand, objects can belong to different semantic categories at the same time (i.e., a baby boy belongs to the categories PERSON, CHILD, BOY, HUMAN, etc.), which, according to Brown (Brown, 1958) could all be valid alternatives for naming that object. On the other hand, the three different levels within semantic categorization[2] identified by Rosch et al. (Rosch et al., 1976) can all be valid alternatives for naming the same object as well. Although the basic-level categories are considered to be the most natural terms for speakers when referring to objects (Hajibayova, 2013; Jolicoeur et al., 1984; Rosch et al., 1976), these are not universal categories since they are restricted by perceptive, cognitive and environmental factors that can result in lexical variation (Berlin, 2014; Graf et al., 2016; Malt, 1995; Wierzbicka, 1996).

While the task of Object Naming has been studied in both Language and Vision and Psycholinguistics, and it is related to Object Recognition tasks in Computer Vision, each field has a different approach to the task:

Within the field of Language and Vision, datasets typically collect free and natural referential utterances[3] produced by annotators for a given real-world image. Some relevant datasets are RefCOCO

[1] Available at https://huggingface.co/datasets/projecte-aina/cat_manynames

[2] The superordinate level (i.e., animal), the basic level (i.e., dog), and the subordinate level (i.e., Chihuahua)

[3] In semantics, a referring expression is a piece of language (typically a noun phrase) used with a particular referent in mind that refers to something or someone, or a clearly delimited collection of things or people (Hurford et al., 2007).

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 31–36
November 20, 2022. ©2022 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_004

(and its newer variant RefCOCO+) (Yu et al., 2016), Flickr30k Entities (Plummer et al., 2015), and VisualGenome (Krishna et al., 2017). Although naming occurs within those datasets, it is not normally marked up and linked to its corresponding image regions.

The task of picture naming constitutes an important experimental paradigm on research in Cognitive Science and Psycholinguistics, and has been traditionally used to assess language impairments and difficulties recalling general knowledge from semantic memory (Snodgrass and Vanderwart, 1980). Subjects reach a high agreement in this task, but it must be taken into account that participants are normally shown line drawing pictures that depict a prototypical category rather than real-world images that show objects in a context.

In Computer Vision, the task of Object Recognition identifies and classifies objects into several different categories (Russakovsky et al., 2015). Nevertheless, current recognition benchmarks use labels and images from ImageNet (Deng et al., 2009) that assume a single ground-truth label, ignoring linguistic variation.

As we can see, the task of Object Naming is addressed differently in Cognitive Science, Language and Vision and Computer Vision, but it would highly benefit from bringing together the particularities of each field so as to generate and provide quality resources.

2.2 The ManyNames Dataset

The ManyNames dataset (Silberer et al., 2020a) provides up to 36 crowd-sourced names for 25K object instances extracted from VisualGenome (Krishna et al., 2017). Unlike other Language and Vision datasets, it focuses on Object Naming rather than collecting complete utterances. Data collection was inspired by the picture naming norms developed in Psycholinguistics (Snodgrass and Vanderwart, 1980) but using real-world images of objects in a visual context, making it suitable for analysis and modeling of object naming, as well as for research in Language and Vision.

Images were selected from seven domains[4] (ANIMALS_PLANTS, BUILDINGS, CLOTHING, FOOD, HOME, PEOPLE, VEHICLES) by defining 52 synsets from VisualGenome in order to collect instances from different taxonomic levels. Instances were sampled depending on the size of the number of names obtained per synset in order to balance the collection. The annotations were collected by setting a crowdsourcing elicitation task on Amazon Mechanical Turk (AMT). The procedure required several annotation rounds, in which problematic cases such as unclear bounding boxes or occluded objects were discarded. Because of noise in the data, a second version of ManyNames (MN v2) was released (Silberer et al., 2020b), which is a verified dataset that contains consistent response sets with adequate responses that refer to the same object only. The resulting dataset contained substantial variation (2.2 names per object on average in MN v2). ANIMALS_PLANTS obtained the highest agreement, whereas PEOPLE reached a particularly low agreement. The analysis performed on the Bottom-Up model (Anderson et al., 2018) using the ManyNames dataset (Silberer et al., 2020b) showed that single-label data underestimated model effectiveness against multi-label data, obtaining a lower accuracy. This demonstrates that, compared to single-label resources for Object Naming, the ManyNames dataset provides a more accurate picture of human naming preferences by taking into account linguistic variation.

3 A New Dataset for Object Naming in Catalan

The main motivations for using the ManyNames dataset as source data are (i) its consideration of linguistic variation in Object Naming, which is widely ignored up to now in Computer Vision, and (ii) the better accuracy that has shown to perform against single-label datasets in Language and Vision modelling. In order to obtain a Catalan version of ManyNames, we decided to automatically translate all the annotations in the original English dataset to Catalan using a state-of-the-art Machine Translation (MT) tool. To assess the quality of the resulting resource, we collected real human annotations for a subset of the dataset, consisting of around 1K images. Although the size of the manually annotated subset may seem small, it can be considered standard for a test set with the purpose of evaluating the quality of automatic annotations. Table 1 shows an overview of the columns contained in the CAT ManyNames dataset.

[4] All domains are based on McRae's feature norms (McRae et al., 2005) except PEOPLE, which was considered to be salient due to its prominence for humans.

Column	Type	Description
responses	dict	Correct responses and their counts
topname	str	The most frequent name of the object
domain	str	The ManyNames domain of the object
incorrect[5]	dict	Incorrect responses and their counts
singletons[6]	dict	All responses which were given only once
total_responses	int	Sum count of correct responses
split	str	Use of the images in training, test and validation
vg_object_id	int	The VisualGenome id of the object
vg_image_id	int	The VisualGenome id of the image
topname_agreement[7]	int	Top name responses divided by total responses
jaccard_similarity[8]	int	Jaccard similarity index of the responses column
raw_responses[9]	dict	Uncorrected responses in the human annotated test set

Table 1: Description of the columns in the CAT ManyNames

3.1 Translated Annotations

Two different neural MT systems were considered before carrying out the translation of the ManyNames dataset: SoftCatalà and Google Translate.

Softcatalà is an open-source initiative [10] that, among other free NLP tools, offers automatic translation services between Catalan and several languages based on neural network technology (Mas, 2021). The popular Google Translate engine, which provides translation services between more than 100 language pairs (Caswell and Liang, 2020), was also considered.

Given the lack of linguistic context in the annotations to be translated (which were, in most cases, a single word), sense disambiguation was a major linguistic issue that needed to be solved before carrying out the automatic translation. Since no current MT system is yet able to take advantage of images as context [11], ad-hoc linguistic contexts were automatically inserted in each input string in order to compensate for this. The linguistic patterns were added using regular expressions depending on the domain. For example, in the domain HOME, the following pattern was used: "I bought a/an [word] for my home." .

Once the linguistic contexts were added, the resulting sentences from the training split of the data were translated with both SoftCatalà and Google translate. In order to evaluate which system performed a better translation, a random sample of 500 sentences out of the total translated sentences was collected and its quality was manually evaluated. 403 sentences out of 500 had an identical translation in both systems, but in 74 cases Google Translate got a more accurate translation than SoftCatalà (which only surpassed Google Translate in 23 examples), probably due to having been trained with larger amounts of data. As a result, Google Translate was considered as a better option for performing the automatic translation of the dataset. The linguistic patterns added in order to disambiguate were removed after the translation of the whole dataset, and repeated words, as well as their counts, were merged.

3.2 Manual Annotation of the Test Set

In order to gather as many manual annotations as possible for the test set, an annotation campaign was launched for a subset of 1,072 images. For this, we used 22 different Google Forms[12], each containing 50 images[13]. Participants were asked to fill one of the Google forms (picked at random) and name the object, animal or person inside the bounding box with the first name that came to their mind. Demographic information about participants was collected during the survey, such as age, gender and region of origin. Statistics show that they were quite balanced in terms of age and gender, but in terms of geographical variation, the Central Catalan dialect was largely over-represented. At the end of the campaign, a total of 220 native Catalan speakers had participated, gathering a total of 10,072 annotations, corresponding to 10 annota-

[10] Visit the following link for further information: https://github.com/Softcatala/nmt-softcatala

[11] Please note that in order to carry out the automatic translation, images were not considered

[12] Among the main reasons to use Google forms are its simplicity of use and the possibility to fill in surveys from a mobile device.

[13] Except the last one, which contained 22 images.

tions per image[14].

Post-processing steps for the human annotations included spellchecking the responses. After this step, possible erroneous responses were filtered out by comparing the corrected responses to the *incorrect* translated column of the ManyNames dataset and were also manually revised. In the process, possible offensive and/or inadequate content were also eliminated. Counts were added once the filtering process was finished. The resulting manually annotated subset has been published with an open license[15].

4 Analysis and Discussion

The purpose of the analysis was, on the one hand, to assess the quality of the automatic translation in the subset that had been human annotated by comparing both the translation and the human annotations, and on the other, to explore possible differences in lexical choices based on cultural biases. To this end, the accuracy of the top name, the degree of variation per image, the average number of different responses per image and the agreement on the top name were computed for both test sets.

The most immediate measure to evaluate the quality of the translated test set was to compute the accuracy of the translation of the most frequent response per image (aka the top name) by comparing it with the corresponding top name in the human annotated set. This accuracy only reached 67,91%, which is a clear indication of how different both resources are.

Another interesting metric to be computed was the degree of lexical variation in the two sets. Despite the difference in the number of annotations (36 for the translated vs 10 for the human annotated), the average number of types in the translated test set was 2.1 responses per image, whereas in the human annotated test set, it was 3.1[16], showing greater lexical variation in the human annotated test set. To account for this clear divergence, we could hypothesize that often two different names get conflated into one in the translation process. However, the ratio of the translated dataset (2.1) is very close to the 2.2 names per object on average in the original ManyNames dataset.

A related metric that was also applied is agreement on the top name per image, which is computed by dividing the number of responses for the top name by the number of total responses. Since more variation is observed in the human annotated set, we expect a higher agreement in the translated set. Indeed, the median is higher in the translated data (0.93) than in the human annotated data (0.7).

A qualitative analysis was performed by sorting both test sets by domain and top name and manually inspecting them to spot divergent cases of translation between English and Catalan. Several findings account for the observed richer lexical variation in Catalan: it was found that Catalan speakers tended to choose a subordinate name (*portaveu, esportista, tennista, etc.*) rather than a taxonomic name (*dona, noi, noia, etc.*) in the PEOPLE domain, the exception being images that involved specific terminology of an activity or a sport not specific to the Catalan culture, i.e. baseball or skateboarding. In those cases, Catalan speakers tended to choose the basic level (*jugador, noi*) rather than the subordinate level (*batedor, patinador*). Certain domains, such as CLOTHING showed Catalan to be more specific than English (which had repercussions in the translation). For example, *jacket* can be translated as *americana* or *jaqueta*, depending on the formality of the event. In addition, Catalan speakers may opt for the use of a diminutive (*trenet* vs *tren*), but this is a lexical option that English speakers do not have.

Our analyses show major divergences between the automatically translated dataset and the manually annotated subset, both in terms of degree of internal lexical variation and accuracy of the translated top names. Manual inspection of the results further confirms that these divergences can be attributed to linguistic and even cultural differences. Automatic translation of language resources from well-resourced languages to less-resourced ones is a common practice in NLP and related fields. Our results show that linguistic and cultural differences may affect the quality of automatically translated resources, such as the one presented here.

5 Conclusions

In this paper, we have presented a new dataset for the task of Object Naming in Catalan, namely CAT ManyNames. The new resource is the result of

[14] Time constraints prevented us from gathering more annotations per image, but for the purposes of the present exercise, 10 annotations looks like an acceptable number

[15] Available at https://huggingface.co/datasets/projecte-aina/cat_manynames

[16] As for the types by domain, the human annotated test set has more types in all domains except in FOOD and CLOTHING, where both test sets have the same number of types.

the machine translation of the English ManyNames dataset, with some pre- and post-processing steps. It also includes a subset of 1,072 images which has been entirely human annotated with 10 annotations per image. The comparison between the translated and the human annotated subsets reveals cultural-based divergences in lexical choices that can affect the quality of the machine-translated resource. Our results shows potential weaknesses in resources built up by translating annotations, particularly in the Language and Vision field, where context is provided by the image and thus is not available to the machine translation system. Since current literature on Object Naming within the Language and Vision field is scarce, these findings could serve as a starting point for research on cross-lingual Object Naming, and on the impact of automatic translation in the annotatation of multilingual resources.

6 Acknowledgements

This research was supported by the Generalitat de Catalunya through the AINA project. We thank our colleagues in the Text Mining Unit at the Barcelona Supercomputing Center, who provided insight and expertise that greatly assisted the research. We would also like to thank the anonymous reviewers for their insightful comments.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Brent Berlin. 2014. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*, volume 185. Princeton University Press.

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

Isaac Caswell and Bowen Liang. 2020. Recent advances in google translate.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *CogSci*.

Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science*, 39(5):676–687.

James R Hurford, Brendan Heasley, and Michael B Smith. 2007. *Semantics: a coursebook*. Cambridge university press.

Pierre Jolicoeur, Mark A Gluck, and Stephen M Kosslyn. 1984. Pictures and names: Making the connection. *Cognitive psychology*, 16(2):243–275.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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(1):32–73.

Barbara C Malt. 1995. Category coherence in cross-cultural perspective. *Cognitive psychology*, 29(2):85–148.

Jordi Mas. 2021. Resum de l'any 2020 a softcatalà.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

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.

Eleanor Rosch, Carol Simpson, and R Scott Miller. 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, 2(4):491.

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(3):211–252.

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 *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905.

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.

Anna Wierzbicka. 1996. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.