Which Languages do People Speak on Flickr? A Language and Geo-Location Study of the YFCC100m Dataset

Alireza Koochali [1, 2] Sebastian Kalkowski [1, 2] Andreas Dengel [1, 2]
Damian Borth [2] Christian Schulze [2]

[1]University of Kaiserslautern, Germany [2]German Research Center for Artificial Intelligence (DFKI), Germany
{alireza.koochali, sebastian.kalkowski, andreas.dengel, damian.borth, christian.schulze}@dfki.de

ABSTRACT

Recently, the *Yahoo Flickr Creative Commons 100 Million* (YFCC100m) dataset was introduced to the computer vision and multimedia research community. This dataset consists of millions of images and videos spread over the globe. This geo-distribution hints at a potentially large set of different languages being used in titles, descriptions, and tags of these images and videos. Since the YFCC100m metadata does not provide any information about the languages used in the dataset, this paper presents the first analysis of this kind. The language and geo-location characteristics of the YFCC100m dataset is described by providing (a) an overview of used languages, (b) language to country associations, and (c) second language usage in this dataset. Being able to know the language spoken in titles, descriptions, and tags, users of the dataset can make language specific decisions to select subsets of images for, e.g., proper training of classifiers or analyze user behavior specific to their spoken language. Also, this language information is essential for further linguistic studies on the metadata of the YFCC100m dataset.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

yfcc100m; flickr; geo-location; language detection; nlp

1. INTRODUCTION

Recently, the training of novel machine learning approaches such as Convolutional Neural Networks (CNN) [16] on top of large scale datasets such as ImageNet [9] or MS Common Objects in Context (COCO) [17] has turned out to improve classification and detection performance significantly. To move further into this direction, the *Yahoo Flickr Creative Commons 100 Million (YFCC100m)* dataset was released [25, 26]. This image and video collection is not only currently the largest freely available dataset in computer vision and multimedia research, but also provides a rich repository of metadata associated with each image and video. This content can be used either to provide tailored access to the dataset or to retrieve specific sub-sets from the dataset for special purposes [14]. The YFCC100m has already been used as source for training classifiers [19] and for visual recognition tasks and multimedia challenges [2, 7].

MMCommons'16, October 16 2016, Amsterdam, Netherlands

ISBN 978-1-4503-4515-6/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983554.2983560

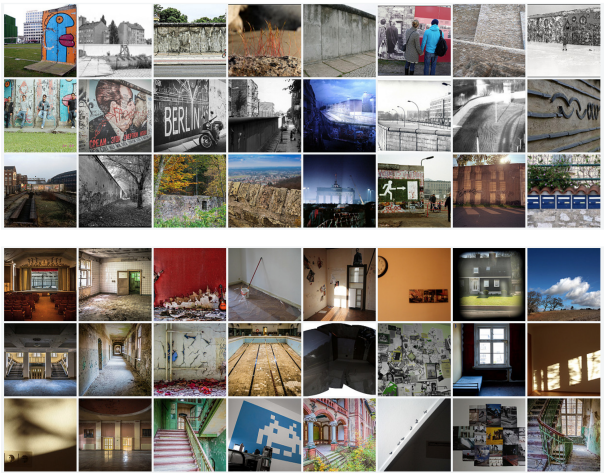

Figure 1: Illustration of language specific differences in the YFCC100m dataset. **Top:** Images depicting a "Mauer", the German word for an *outdoor wall*. **Bottom:** Images depicting a "Wand", the German word for an *indoor wall*. Such language specific differences can have a significant impact if not taken into account.

To fully embrace the potential of this global dataset, we have to consider language as a major but missing parameter for extended utilization of the dataset. Different languages differ in their understanding and embedding of semantics as illustrated in Fig. 1: the German language makes a difference between an *outdoor wall* ("Mauer") and an *indoor wall* ("Wand"), while in English there is one common term ("wall") describing both. Images retrieved by the corresponding German words differ significantly in their visual appearance. For example the extension of the Visual Sentiment Ontology (VSO) [4] to Multilingual Visual Sentiment Ontology (MVSO) [13] to cover cultural differences in the visual appearance of concepts was done entirely by using query terms in different languages. Vice versa, the MS COCO dataset is specifically designed to represent objects

by pictographs to work around the language barrier [17]. Although the research community has already provided access and browsing capabilities to the dataset [14], language as an important dimension of the metadata was until now not investigated within the YFCC100m dataset.

This paper presents the first analysis of language on the YFCC100m dataset. It runs Natural Language Processing (NLP) algorithms on all meta-data including titles, descriptions, and tags of each item and provides (i) global, (ii) user, and (iii) item specific information about the languages used in the dataset. Additionally, it links this information to the geographic distribution of the images and videos in the dataset and therefore allows to understand the interplay between language and geo-location as observed on the Flickr platform. Finally, all data will be made publicly available to enrich the YFCC100m dataset including an online tool to interactively investigate language-specific properties of the dataset[1].

2. RELATED WORK

This section describes related work with respect to the dataset itself and related work about language analysis made on Flickr and work dealing with geo-location data on Flickr.

Geo-Location on Flickr.

Popular services for sharing images and videos on the internet, including Twitter, YouTube and Flickr provide support for geo-locations [23]. Research with geo-data on Flickr has a long tradition [1, 15, 12, 11]. Either by analyzing geo-data [1] or estimating geo-locations from textual [12], visual [15, 12, 11], or audio content [10].

In the case of the YFCC100m, previous research reported that this additional location information is provided surprisingly often namely in 48.3% of all images and videos [14]. This is presumably due to newer camera and phone models with GPS support. In this context dedicated challenges exist, where the YFCC100m dataset is used. Among those are the *MediaEval Placing Task* [7], which aims for an automated estimation of image location, from its pixel- and meta-data. For this task, the *YLI-GEO* [2] dataset, a subset of the geo-tagged images and videos from the YFCC100m dataset serves as ground truth for training and evaluation. A similar objective has the *Yahoo-Flickr Event Summarization Challenge* as part of the *ACM MM 2015 Grand Challenge*. Here, the goal is to automatically detect and identify structures in the YFCC100m collection, which allow event detection, description, and summarization of detected events [5], using geo-information as one of the possible information sources.

Languages on Flickr.

Natural Language Processing (NLP) is a subcategory of text mining and information extraction focused on human language understanding. In recent decades, successful applications of NLP established in the domains of language detection, automatic translation, as well as spelling and grammar correction [3]. Another area of NLP is Named Entity Recognition (NER), aiming to extract entities like person-, organization-, and location names from unstructured text. For a survey about current methods in NER please refer to [18].

[1]www.yfcc100m.org

Table 1: This table illustrates the distribution of detected languages in titles, descriptions, and tags of the YFCC100m datasets. It can be observed that the precision of language detection differs with the type of input i.e. it is more difficult to infer a language from a short title than from a multi-sentence description.

Titles		
w/o title	3,835,258	3.8%
generic titles [1]	25,971,801	26%
non-generic titles	70,192,941	70.1%
- language not detected	39,367,031	39.3%
- with one language	30,715,239	30.7%
- with two languages	109,170	0.1%
- with three languages	1,501	0.001%
Descriptions		
w/o description	68,277,216	68.3%
with description	31,722,784	31.7%
- language not detected	7,024,604	7%
- with one language	24,257,091	24.25%
- with two languages	416,084	0.4%
- with three languages	25,005	0.02%
Tags		
w/o tag	31,028,877	31%
at least 1 tag	68,971,123	69%
- language not detected	26,145,788	26.1%
- with one language	41,539,461	41.5%
- with two languages	1,256,732	1.2%
- with three languages	29,142	0.02%

[1] e.g. "IMG_012345" or "DSC_12061999", lower bound.

With respect to Flickr, most of the studies regarding language processing focus on tags only. One study shows the importance of language detection on image tags on Flickr for understanding the semantics of data to generate an index [20]. Another study discusses, how language models of textual data can be employed, to determine the image-location. [29, 22] or even use language models trained by Flickr data to geo-referencing of other sources like Wikipedia pages [8]. In the area of NER, people proposed ontology based detection algorithms with a subsumption-based model on the tags of 5 million images from Flickr [21]. Other examples of applying NER on Flickr images include linking knowledge-base entities with their corresponding entities in images, i.e., to populate an available knowledge base with photos of Named Entities [24].

In contrast to previous works, this paper focuses on the investigation of spoken languages in the YFCC100m dataset and its relation to geographic locations. This analysis provides unique insights, which are currently not available for the YFCC100m dataset but are of high value to further research and proper usage of the dataset.

3. YFCC100M DATASET

Flickr is a popular online platform for sharing images and videos. This user generated content is enriched by its users with titles, descriptions, and tags and by the capturing devices with geo-location and EXIF information. Flickr also allows to upload content as Creative Commons, making it

one of the largest repositories of freely available images and videos. The YFCC100m dataset is entirely compiled of such Creative Commons images and videos from the Flickr platform. It can be downloaded free of charge from Yahoo Labs[2] including the available metadata for all 100 million images and videos. Additionally and separately the actual image and video content can then be retrieved from Flickr servers without the use and limitations of the Flickr API.

3.1 Title, Description, Tag Information

Users on Flickr have the option to annotate their images with additional information such as title, description, and tags. Although noisy [27], such optional features provided by the platform allow to refine search parameters and narrow down the retrieval of images. However, not all users utilize these annotation features. It can be seen in Table 1 that although 96% of items have non-empty titles, a large fraction of titles (at least 26%) is auto-generated consisting of non-descriptive strings such as "IMG_012345" or "DSC_12061999", determined by matching a simple Regular Expression. Such generic titles can not be linked to a specific language and therefore are excluded in language analysis. Further, users on Flickr are not very keen on providing descriptions for their images. Only 31.7% of the datasets images have a non-empty description. This is different for tags. Here, for more than 69% of the items users entered at least one tag.

3.2 Geographic Information

Roughly 48.3% of all images and videos in the YFCC100m provide geo-location data. Such information is usually given in form of latitude and longitude values, defining a position on the globe. For over 99% of all geo-tagged images and videos it is possible to determine the country in which they were acquired. Similar mappings have also been done in previous works [14], where also the plausibility and consistency of the geo-locations (if given) have been approved.

Figure 2 visualizes the distribution of the complete set of geo-located images and videos of the YFCC100m dataset over the world on a logarithmic scale This map reveals a strong bias of the dataset towards the USA in particular, but also Brazil, India, Australia, Central Europe, China and Japan show relatively high contributions to the dataset in comparison to other regions. In consequence, those active regions are over-represented in the dataset, while the least active states in turn are comparably underrepresented. By normalizing the absolute numbers of images and videos for each country by the respective numbers of inhabitants, we get a distribution as visualized in Figure 3. Here we see that the ratio of contributed images and videos per person is highest in three major hot-spots: Europe, Japan and the USA. This means, compared to the number of people living in those areas - thus representing their culture - those areas can be expected to provide the most comprehensive and accurate picture of their cultural identity.

The implications of this are two-fold: First, this knowledge is of major concern, when attempts are made to derive social behavior analysis from the YFCC100m dataset.

[2]https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67

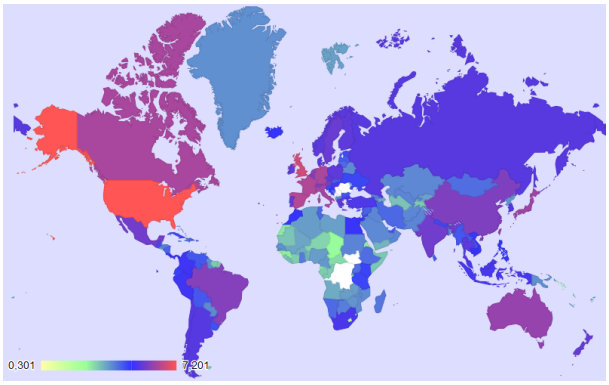

Figure 2: Geographic distribution of all geo-located images and videos in the dataset on a logarithmic scale. We see a bias of the dataset towards certain countries, in particular the USA.

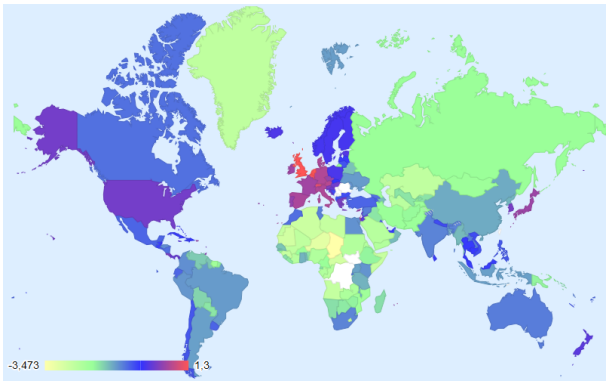

Figure 3: Geographic distribution, normalized by the approximate population for each resp. country, on a logarithmic scale. Japan, Europe and the USA turn out to be hotspots concerning the cultural representativeness within the dataset.Note that the few white areas represent countries, which had to be excluded from the visualization due to map data inconsistencies between the geo-coding and visualization

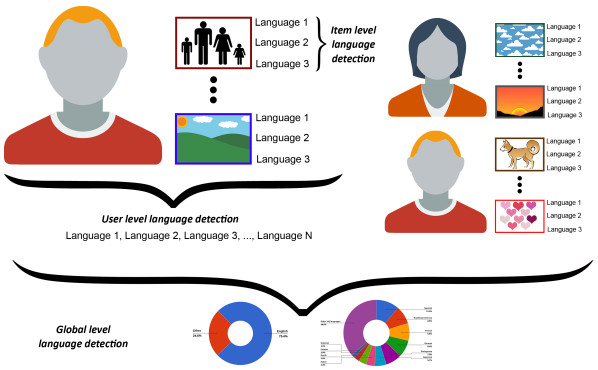

Figure 4: Illustration of different levels of language analysis. First, image level analysis on title, description, and tags. Second, user level analysis to list top languages for each user. And thirdly, a global aggregation is done to provide language findings for the entire YFCC100m dataset.

4. LANGUAGE ANALYSIS

As mentioned, the YFCC100m dataset is exclusively curated from user generated content, including various textual information such as, e.g., titles, descriptions and tags. Its global spread and enormous amount of textual data makes it an excellent choice for linguistic studies and natural language processing. This section provides a language analysis of the dataset on different levels of granularity. As illustrated in Fig 4, first, for each image a language detection is performed on the image's title, description, and tags to list the top 3 languages. Second, a user language analysis is done to list top detected languages for each user in the dataset. Third, a global analysis is done to aggregate item or user specific information about the use of language.

4.1 Language detection

The first analysis step on the dataset is language detection. Information about the specific language (e.g. English, French, . . .) is essential for further natural language processing steps and linguistics studies, e.g., Named Entity recognition, Part of Speech tagging etc. It also provides a natural way to aggregate users, and to learn about language use within the YFCC100m dataset.

Currently, most of the language identification approaches rely on character n-grams or byte n-grams comparing n-gram profiles, or using various machine learning classifiers [6]. As one of the most sophisticated NLP frameworks we employ Bloomberg's NLP<GO>[3]. NLP<GO> is an extensible, high-performance, open-source library designed for building and running complex Natural Language Processing (NLP) applications. It includes state-of-the-art algorithms for the most common NLP tasks. For language detection, NLP<GO> uses Google's Compact Language Detector 2 (CLD2). CLD2 is a Naive Bayesian classifier, which looks at quad-grams and examines them against a very large reference token table to calculate scores. The reference table has been created from the training corpus, which is manually constructed from selected web pages for each language, and augmented by careful automated scraping of over 100M additional web pages. For mixed-language input, CLD2 returns a ranked list of the top three languages found and their approximate percentages of the total text bytes.

For our analysis, we used the original algorithm without any modification, assuming that the accuracy of this state of the art tool also applies for our use case. Information about reliability and accuracy of CLD2 can be taken from the project website[4].

For each image, we detect the language(s) within its title, description and concatenation of all tags separately. Potentially, we can detect up to 9 different languages per image as CLD2 can return up to 3 languages for each input. However, in practice, each image usually has at most 3 different languages. Statistics yield that globally about 40% of images have more than one language. This raises the question weather all detected languages for one item (title, description, tags) are used equally, or the input is biased towards one language. For example imagine a user traveling to New York and annotating its title as "Big Apple" – an English word – but its description would be written entirely in French whereas the tags would be a mixture of English and French words. To take such behavior into account, we utilized borda count [28] as a method for fusing the ranks of detected languages from each annotation type into a single rank for each image or user respectively. Based on the borda count principle, a score is assigned to each language, proportional to its rank on the detection list. As the list length, at most is 3, the first language scored with 3, the second language scored 2 and the third language scored one. The final list is then ranked by the sum of each languages scores from the different inputs.

Results on image level show great diversity between detected languages. We have found 172 different languages in total, from common languages to rare dialects, artificial languages and extinct languages. We used borda count across all images to rank the occurrence of languages throughout the entire dataset annotations. Fig. 5 shows how English dominates the dataset and Fig 6 shows the top 10 most occurring languages - excluding English - over the dataset. It can be seen that languages such as Spanish, Traditional Chinese, French, German, Portuguese, and Japanese are dominant as already indicated in the geographic distribution of the images and videos of the dataset. Moving up the aggregation level, images can also be grouped by the responsible users unique ID (assuming that a user on Flickr represents one individual person in the real world). Consequently, we can determine a set of all languages utilized by a user for his/her images alongside their borda count. Figure 7 demonstrates that more than half of the users are annotating with

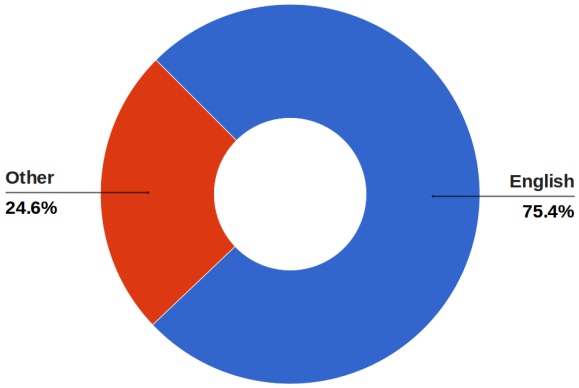

Figure 5: Distribution of English vs other languages among all the images in the YFCC100m that at least have one language detected (approximatley 62% of all images).

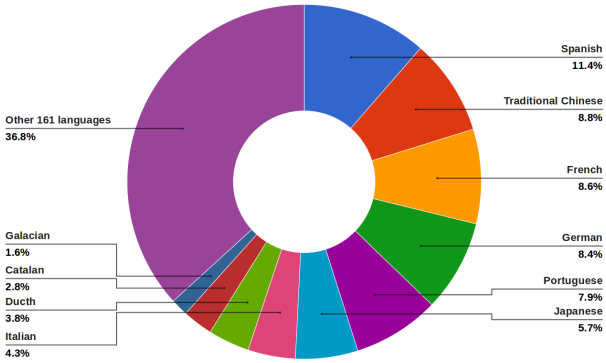

Figure 6: Distribution of top ten languges excluding English. Top three languages are Spanish, Traditional Chinese and French.

[3]https://github.com/bloomberg/nlpgo
[4]https://github.com/CLD2Owners/cld2

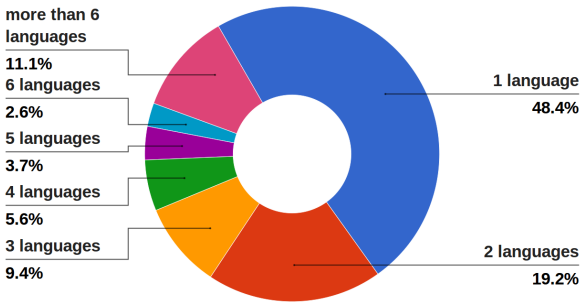

Figure 7: Number of languages detected for each user.

Table 2: Top ten languages used as the first (second) language alongside the number of their users.

	First Lang.	Users	Second Lang.	Users
1.	English	344788	English	40306
2.	Spanish	23450	French	10925
3.	Portuguese	13045	Spanish	10580
4.	German	11145	German	8393
5.	French	9984	Scots	7530
6.	Trad. Chinese	8132	Portuguese	6241
7.	Italian	7916	Italian	5767
8.	Japanese	3308	Latin	5201
9.	Danish	2993	Japanese	3909
10.	Chinese	2763	Indonesian	3852

more than one language, which is showing the great potential for linguistics studies. Surprisingly, roughly 9.2% of users are even using more than seven languages within their images and videos. Manual inspection yielded that most of these accounts are shared accounts, i.e., they violate our previous assumption of one account to one person associations. To understand more about languages, which are spoken as first or second language on YFCC100m, we define the language with the highest borda score in a user's set of languages to be his first language and the language ranked second as his second language. Table 2 lists top ten first languages and second languages on YFCC100m in conjuction with the number of users sharing them.

Previously, we perceived that the English language dominated the dataset and by considering the user-level result, we also perceive that it also has the highest number of first language speakers. This brings up the question, which language community is the most active one between all communities. For measuring this, we define the "Active Community Factor" (ACF) as follows:

$$\frac{Total\ borda\ count\ of\ language\ L}{Number\ of\ users\ having\ L\ as\ first\ language} \quad (1)$$

Having the "Active Community Factor" for the top ten languages, we can see that English doesn't take the first place anymore, as shown on Table 3. If we calculate the active community factor on the whole dataset, the communities of artificial languages like Klingon and Pig Latin are getting even a higher score than English (obviously because there is only a small number of people who are however forming a very active community).

Finally, we attempted to push one step further and find the relation between first and second languages. The results are depicted in Fig. 8, which shows the top 30 relations between first and second languages comprehensively. Surprisingly, most of people using English for annotations do not use any other language in addition. This could be another reason why the English language dominates the dataset. It also indicates that most of the English speakers either do not have a (sufficiently proficient) second language or they do not see any necessity to use their second language in social media. This contradicts the behavior of most of the speakers of other first languages. They most often use English as their second language and most probably do this for reaching a larger audience and might have come to the consensus to use English as an international standard.

Table 3: Top ten languages reordered based on their Active Community Factor (ACF)

Language	first language users	sum borda count	ACF
Japanese	3308	2727932	824.64
Dutch	2500	1812932	725.17
Traditional Chinese	8132	4212493	518.01
Catalan	2710	1323356	488.32
English	344788	146903404	426.06
French	9984	4128703	413.53
German	11145	4010544	359.85
Portuguese	13045	3787845	290.36
Italian	7916	2073709	261.96
Spanish	23450	5467660	233.16

Table 4: Top ten detected Named Entities and their count alongside the typical forms for each Named Entity (more than 100 times) and the count of each form separately

Named Entity	Typical form	Total Count
France	france(517104)	517137
California	california(432375) _california(408)	432901
Chicago	chicago(299917)	299999
New york	new york(279260) new-york(492) new york(145) new_york(108)	280182
San francisco	san francisco(244489) sanfrancisco(29125) san_francisco(330) san_francisco(170)	274323
Spain	spain(265776)	265782
Washington	washington(255241)	255290
Mexico	mexico(237135)	237159
China	china(198690)	198743
Singapore	singapore(186198)	186204

4.2 Named Entity Recognition

If a user took a picture of the Times Square, we would expect to see the words "Times Square" somewhere in the title, description or among tags. It is obvious that Named Entities have a very strong relation to image content. Therefore, recognition of Named Entities in textual data can provide a set of words that describe the contents of the dataset precisely. NLP<GO> uses the MIT Information Extrac-

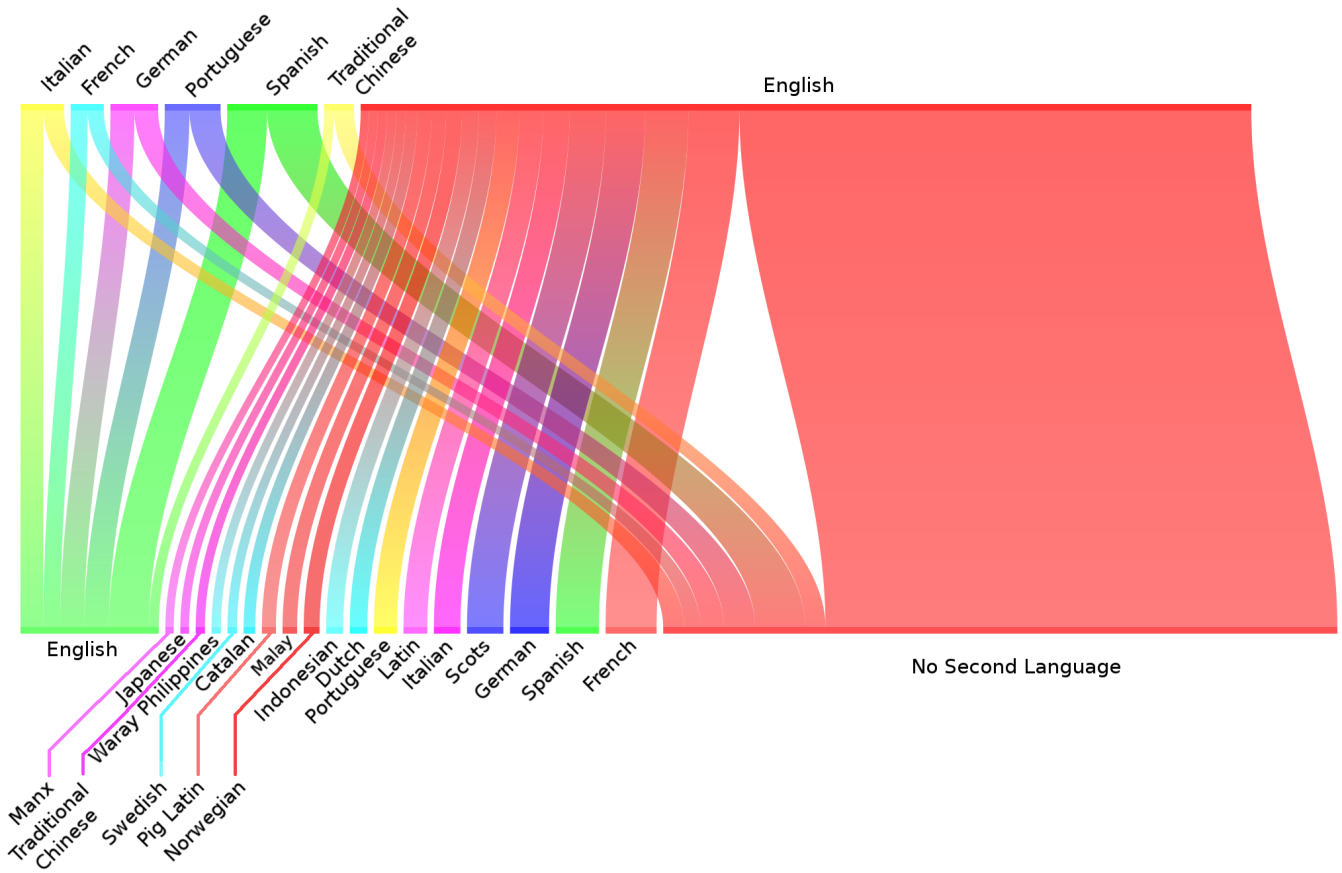

Figure 8: Relation between first and second languages of users. On the top the first languages are shown and on the bottom the second languages (including "no second languages").

tion (MITIE) library for NER tasks. MITIE is an open source information extraction tool, developed by the MIT NLP lab, which is built using state-of-the-art statistical machine learning tools. It comes with trained models for English and Spanish. The English Named Entity recognition model is trained based on data from the English Gigaword News corpus, the CoNLL 2003 Named Entity Recognition Task, and ACE data.

We detected Named Entities using NLP<GO> and counted the number of their occurrences in the dataset. The results of NER show that locations are the most used type of Named Entities in the dataset. Table 4 demonstrates the top ten Named Entities and their number of occurrences. Additionally, it provides the most common forms that each of them typically appeared in (if it at least appeared more than 100 times) with their counts.

5. LANGUAGE GEO INTERPLAY

The geographic distribution of images and the language distribution within the dataset, were expected to be highly dependent on each other. In fact, as depicted by Fig. 9, most of the time, there is a high correlation between the language used for the textual metadata and the country the image was taken at. Usually official languages score very high in the usage rankings for the respective countries. The other direction also holds most often: languages are mostly used in the respective countries where they are an official language. However, the role of the English language is outstanding in comparison to other languages in the dataset, since its contribution is quite noticeable in most countries. In many countries, English is even used approximately as often as the native languages. In some cases, like France, Italy or Japan, English is even used more often for annotations, than the respective official languages. The reasons for this might be twofold: First, English may be used by Flickr users all over the world, besides their native language, to make images and videos retrievable world-wide by the Flickr search mechanism. The Flickr community – in other words – probably has agreed to the use of English as a common language for international communication. A second factor, leveraging the usage of English within a foreign country borders might be native English-speakers traveling to the respective countries. As shown before, since the proportion of both, images from the United States and Flickr users with English as their most used language, one can assume that the proportion of native English-speakers is comparably high, also leading to an over-representation of English-speaking travelers. To diminish the influence of highly active Flickr contributors to the country-language relations, a similar evaluation has also been done by counting distinct users instead of images and videos for each country-language pair. If we break down the occurrences of country-language pairs to the number of distinct users, we get a relation as depicted in Figure 10. In this visualization, the impact of the USA as a main location and English as a major language declines in comparison to other county-language pairs. However, the previously described tendency of users, tagging and describing their images in

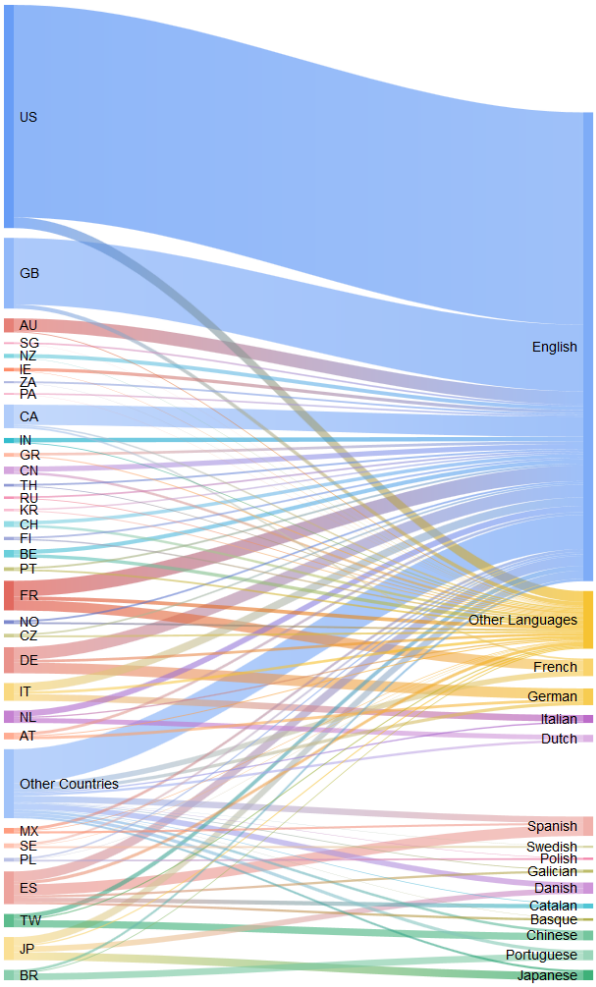

Figure 9: The top 50 combinations of countries and languages on an image-level. The edge-weights are proportional to the number of images taken in the country and using the respective language within its annotations. The pre-eminent role of the English language becomes apparent.

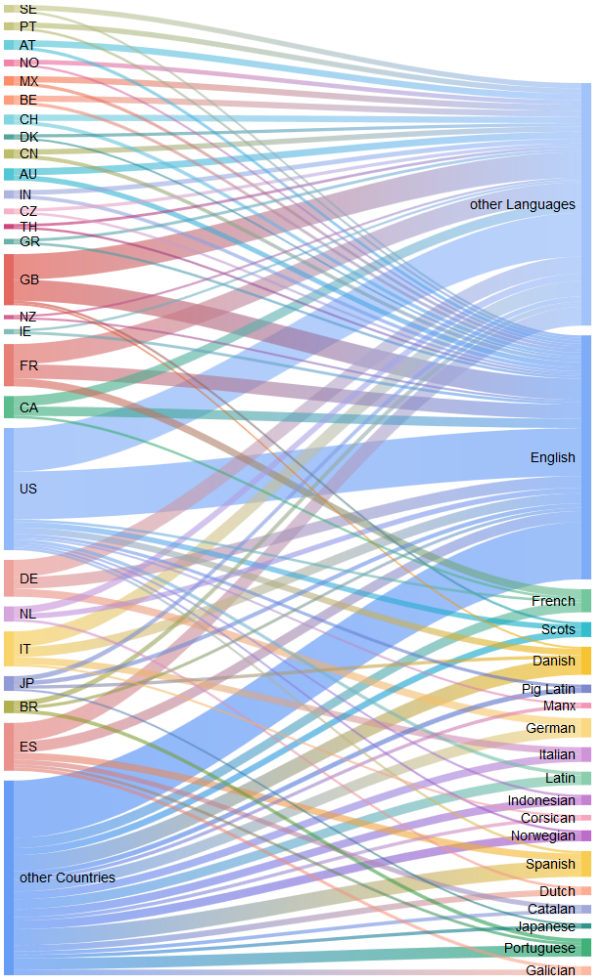

Figure 10: The top 50 combinations of countries and languages, similar to Figure 9 but on a user-level. The edge-weights are proportional to the number of unique Flickr users that contributed at least one image with the respective country-language-pair. The impact of Flickr accounts using only English for all their images declines in comparison to multi-lingual Flickr users.

English approximately as often as in their local language, is still apparent.

6. CONCLUSION

According to our analysis, the YFCC100m dataset – although dominated by the English language – is a very rich and diverse dataset with respect to language. This richness of languages represents the global coverage of the dataset and provides interesting insights about its usage such as the distribution of first and second languages spoken on the YFCC100m. The presented study further illustrates an interesting interplay of geographic locations to languages, indicating that very often users annotate images with languages that are not the most common for the country they are taken at (most often English). In general, language – besides geo-locations – has a crucial influence on how the dataset is used and especially how images and videos from the dataset can be retrieved.

7. ACKNOWLEDGMENTS

This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002).

References

[1] S. Ahern et al. "World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-referenced Collections". In: *ACM/IEEE Conf. on Digital libraries*. 2007.

[2] J. Bernd et al. "The YLI-MED Corpus: Characteristics, Procedures, and Plans". In: (2015). arXiv: 1503.04250 [hep-th].

[3] Anuja Bharate and Devendra Gadekar. "Survey Paper on Natural Language Processing". In: *International Journal of Inventions in Engineering and Science Technology* (2015).

[4] D. Borth et al. "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs". In: *ACM Int. Conf. on Multimedia (ACM MM)*. 2013.

[5] *Call for Multimedia Grand Challenge Solutions*. 2015. URL: http://www.acmmm.org/2015/call-for-contributions/multimedia-grand-challenges/.

[6] William B Cavnar, John M Trenkle, et al. "N-gram-based text categorization". In: *Ann Arbor MI* 48113.2 (1994), pp. 161–175.

[7] J. Choi et al. "The placing task: A large-scale geo-estimation challenge for social-media videos and images". In: *ACM Workshop Geotagging and Its Applications in Multimedia*. 2014.

[8] Chris De Rouck et al. "Georeferencing Wikipedia pages using language models from Flickr". In: *Int. Conf. on Semantic Web*. 2011.

[9] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. July 2009, pp. 248–255.

[10] Benjamin Elizalde et al. "City-Identification of Flickr Videos Using Semantic Acoustic Features". In: *IEEE Int. Conf. on Multimedia Big Data*. 2016.

[11] Gerald Friedland, Jaeyoung Choi, and Adam Janin. "Video2gps: a demo of multimodal location estimation on flickr videos". In: *ACM Int. Conf. on Multimedia (ACM MM)*. 2011, pp. 833–834.

[12] A. Gallagher et al. "Geo-location inference from image content and user tags". In: *CVPR Workshops*. IEEE. 2009, pp. 55–62.

[13] Brendan Jou et al. "Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology". In: *ACM Int. Conf. on Multimedia (ACM MM)*. 2015, pp. 159–168.

[14] S. Kalkowski et al. "Real-time Analysis and Visualization of the YFCC100M Dataset". In: *MM COMMOMS Workshop*. 2015.

[15] Evangelos Kalogerakis et al. "Image Sequence Geolocation with Human Travel Priors". In: *IEEE CVPR*. 2009, pp. 253–260.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proc. Advances in Neural Information Processing Systems (NIPS)*. 2012.

[17] Tsung-Yi Lin et al. "Microsoft COCO: Common objects in context". In: *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.

[18] David Nadeau and Satoshi Sekine. "A Survey of Named Entity Recognition and Classification". In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[19] K. Ni et al. "Large-Scale Deep Learning on the YFCC100M Dataset". In: (2015). arXiv: 1502.03409 [hep-th].

[20] A. Rorissa. "A comparative study of Flickr tags and index terms in a general image collection". In: *J. Am. Soc. Inf. Sci.* 61 (2010), pp. 2230–2242.

[21] Patrick Schmitz. "Inducing Ontology from Flickr Tags". In: (2006).

[22] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. "Placing Flickr Photos on a Map". In: *ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*. 2009, pp. 484–491.

[23] C Szongott, Benjamin Henne, G von Voigt, et al. "Big data privacy issues in public social media". In: *IEEE Int. Conf. on Digital Ecosystems Technologies (DEST)*. 2012.

[24] Bilyana Taneva, Mouna Kacimi, and Gerhard Weikum. "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity". In: *ACM Int. Conf on Web Search and Data Mining (WSDM)*. 2010, pp. 431–440.

[25] B. Thomee et al. "The New Data and New Challenges in Multimedia Research". In: (2015). arXiv: 1503.01817 [hep-th].

[26] B. Thomee et al. "YFCC100M: The New Data in Multimedia Research". In: *Comm. ACM* 59.2 (2016).

[27] A. Ulges, D. Borth, and T. Breuel. "Visual Concept Learning from Weakly Labeled Web Videos". In: *Video Search and Mining*. Springer, 2010, pp. 203–232.

[28] Merijn Van Erp and Lambert Schomaker. "Variants of the borda count method for combining ranked classifier hypotheses". In: *Workshop on Frontiers in Handwriting Recognition*. Citeseer. 2000.

[29] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. "Finding Locations of Flickr Resources Using Language Models and Similarity Search". In: *ACM Int. Conf. on Multimedia Retrieval (ICMR)*. 2011, 48:1–48:8.