Categorization in the Wild: Category and Feature Learning across Languages

Lea Frermann (lfrermann@unimelb.edu.au)
School of Computing and Information Systems University of Melbourne

Mirella Lapata (mlap@inf.ed.ac.uk)
Institute of Language, Cognition and Computation
University of Edinburgh

Abstract

Categories such as ANIMAL or FURNITURE play a pivotal role in processing, organizing, and communicating world knowledge. Many theories and computational models of categorization exist, but evaluation has disproportionately focused on artificially simplified learning problems (e.g., by assuming a given set of relevant features or small data sets); and on English native speakers. This paper presents a large-scale computational study of category and feature learning. We approximate the learning environment with natural language text, and scale previous work in three ways: We (1) model the full complexity of the learning process, acquiring learning categories and structured features *jointly*; (2) study the generalizability of categorization models to five diverse languages; and (3) learn categorizations comprising hundreds of concepts and thousands of features. Our experiments show that meaningful representations emerge across languages. We further demonstrate a joint model of category and feature acquisition produces more *relevant* and *coherent* features than simpler models, suggesting it as an exploratory tool to support cross-cultural categorization studies.

Keywords: Categorization; Bayesian modeling; Computational cognitive modeling; Natural language

Introduction

Categories such as ANIMAL or FURNITURE are fundamental cognitive building blocks allowing humans to efficiently represent and communicate the complex world around them. Concepts (e.g., *dog*, *table*) are grouped into categories based on shared properties pertaining, for example, to their `appearance`, `behavior`, or `function`.[1] Categorization underlies other cognitive functions such as perception (Schyns & Oliva, 1999; Goldstone, 2003) or language (Waxman & Markov, 1998; Borovsky & Elman, 2006). Research suggests that categories are not only shaped by the world they represent, but also by the language through which they are communicated (Gopnik & Meltzoff, 1987; Waxman & Markow, 1995). Mental categories exist in every human culture, however their manifestations differ (Malt, 1995; Ji, Zhang, & Nisbett, 2004). The majority of human and computational studies of categorization to date has focused on English speakers (Medin, Unsworth, & Hirschfeld, 2007). We investigate how computational models of categorization extend across languages, and present a data set, model and evaluation framework to this end.

Substantial research interest in categorization has resulted in numerous theories (Nosofsky, 1988; Rosch, 1973; Corter & Gluck, 1992; Murphy & Medin, 1985) which have been thoroughly tested through laboratory experiments as well as computational simulations. Empirical studies are predominantly based on small-scale laboratory experiments, where participants are presented with small sets of often artificial concepts with restricted features (Bornstein & Mash, 2010; Medin & Schaffer, 1978; Kruschke, 1993). This contrasts with category learning *in the wild*, where humans learn from repeated, noisy observations, necessitating to disentangle relevant features from irrelevant ones. Feature and category learning mutually inform one another (Goldstone, Lippa, & Shiffrin, 2001; Schyns & Rodet, 1997). There is also evidence that features are themselves structured to represent the diversity and complexity of the properties exhibited in the world (Ahn, 1998; Spalding & Ross, 2000). This paper investigates the impact of a joint model of categories and their structured features on the quality of representations.

Concept	Natural Language Stimuli	
cat	Les *chats* sont poilus.	猫有尾巴和爪子。
	Cats are carnivores.	Die *Katze* miaut!
dog	الكلب لديه الفراء.	Les *chiens* ont des queues.
	Hunde essen Fleisch.	Look, the *dog* is playing!
apple	I want to eat an *apple*.	*Äpfel* sind rot oder grün.
	苹果在树上生长。	An *apple* contains seeds
kiwi	Can you cut me a *kiwi*?	*Kiwis* sind innen grün.
	كيويس لديها بذور.	Ce *kiwi* est savoureux.

Figure 1: Illustration of model input for five languages for two concepts from categories ANIMAL and FRUIT.

Even though multilingual taxonomy induction has received recent research attention (De Melo & Weikum, 2010), to the best of out knowledge, we present the first large-scale cross-lingual computational study of category and feature learning with a cognitive focus. We compare two cognitively motivated Bayesian models of varying complexity and a word co-occurrence based model. We approximate the complexity of the learning environment with natural language, which has been shown to redundantly encode much of the non-linguistic information in the natural environment (Riordan & Jones, 2011) and influence the emergence of categories (Gopnik & Meltzoff, 1987; Waxman & Markow, 1995). Figure 1 illustrates the input to our models: Following prior work (Fountain & Lapata, 2011; Frermann & Lapata,

[1] We denote (superordinate level) CATEGORIES (ANIMAL or VEHICLE) in small caps; (basic level) *concepts* (*cat* or *car*) in italics, and `feature types` (`function` or `behavior`) in `true type`.

2016), we create language-specific sets of stimuli, each consisting of a mention of target concept (e.g., *apple*), within its local linguistic context (e.g., {contains, seeds}). We consider each stimulus an observation of the concept, i.e., the word referring to the concept is an instance of the concept itself, and its context words are a representation of its features.

Our experiments expose all three models to (1) five diverse languages[2] and (2) rich and noisy natural language stimuli covering hundreds of concepts and thousands of features (Table 1). In sum, the contributions of this paper are:

- The first large-scale, multilingual study of categorization, suggesting a potential of structured Bayesian models to inform cross-cultural categorization studies.
- Evidence that models which learn categories and *structured* features *jointly*, resembling human learning, acquire better representations according to native speakers in all languages.
- A multilingual dataset of categories, concepts and natural language stimuli to support future computational categorization studies across languages.[3]

Computational Framework

Computational models have been used successfully to shed light on a wide variety of cognitive phenomena (Chater, Oaksford, Hahn, & Heit, 2010) including language acquisition (Xu & Tenenbaum, 2007), generalization and reasoning (Griffiths & Tenenbaum, 2006), as well as categorization (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011; Frermann & Lapata, 2016; Kruschke, 1993; Fountain & Lapata, 2011). Here we use computational models to study (1) how categories and their structured representations emerge together from a rich and noisy environment approximated by natural languages; and (2) their generalizability across languages. Our five target differ in both typology and available corpus size (Table 1), allowing us to study the robustness of models of differing complexity to small data. Our study includes three models.

BCF is a **B**ayesian model of **C**ategories and their **F**eatures (Frermann & Lapata, 2015). Given a set of natural language stimuli of concept mentions in local context (Figure 1), it learns categories k as groups of observed *concepts* c, feature types g as clusters of observed features (context words) f, and associations between categories and feature types. Feature types are clusters of features which pertain to distinct properties (e.g., `behavior` or `function`) and are shared across categories. Figure 2 shows example representations learnt by BCF from the English Wikipedia. More formally, we can describe the model through its generative story. We assume a global multinomial distribution over categories $Mult(\theta)$, drawn from a symmetric Dirichlet distribution with hyperparameter α ($Dir(\alpha)$). For each category k, we assume an independent set of multinomial parameters over feature types μ_k, drawn from $Dir(\beta)$, capturing the associations. For each concept type ℓ, we draw a category k^ℓ from $Mult(\theta)$. Finally, for each feature type g, we draw a multinomial distribution over features $Mult(\phi_g)$ from $Dir(\gamma)$. Stimuli d are generated as follows: (1) retrieve the category k^{c_d} of the observed concept c_d; (2) generate a feature type g_d from the category's feature type distribution $Mult(\mu_{k^{c_d}})$; (3) for each context position i, generate feature $f_{d,i}$ from the feature type's distribution $Mult(\phi_{g^d})$. As exact inference is intractable, we employ collapsed Gibbs sampling as described in (Frermann & Lapata, 2015).

BayesCat is a **Bayes**ian **cat**egorization model (Frermann & Lapata, 2016) similar to BCF, however it represents categories through unstructured bags-of-features. As such, the model structure of BayesCat is closely related to topic models (Blei, Ng, & Jordan, 2003). BayesCat learns a global category distribution $Mult(\theta)$, and models each category k as a distribution over concepts $Mult(\phi_k)$ and a separate distribution over features (context words) $Mult(\psi_k)$ each drawn from a separate Dirichlet prior. Stimuli d are generated by (1) drawing a category k_d from $Mult(\theta)$; (2) drawing a concept from $Mult(\phi_{k_d})$; and (3) drawing features f_i from $Mult(\psi_{k_d})$. We derive a hard categorization from BayesCat's soft assignments of concepts to categories by assigning each c to category k, s.th. $argmax_k p(c|k)$. We construct feature types g from BayesCat representations post-hoc, by first representing each feature f as its probability under each category $p(k|f)$, and clustering features into feature types g using k-means. We compute category-feature type associations as $p(g|k) = \sum_{f \in g} \psi_k^f$. We use the collapsed Gibbs sampler of Frermann and Lapata (2016) for inference.

Co-occurrence model We devise a word co-occurrence based baseline to study the benefit of structured Bayesian models over raw text. Each concept c is represented as a vector of co-occurrence counts with features f (context words), capped by a minimum number of required observations, approximating the concept-feature association $assoc(c,f) = \mathcal{N}(c,f)$. We obtained categories by clustering concepts based on their vector representations using k-means. Based on these categories, we obtained feature types by (1) collecting all features associated with at least half the concepts in the category; and (2) clustering these features into feature types using k-means clustering.

Data

Our experiments focused on 491 basic-level concepts, taken from two previous studies of concept representation (McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008), for which our models learn (a) a categorization and (b) structured feature representations. Human-created gold standard categorizations of the concepts into 33 categories

[2] English, German (both Germanic), Arabic (Semitic), Mandarin Chinese (Sinitic), French (Romance)

[3] The data set can be downloaded from `http://frermann.de/multiling_categories/index.html`.

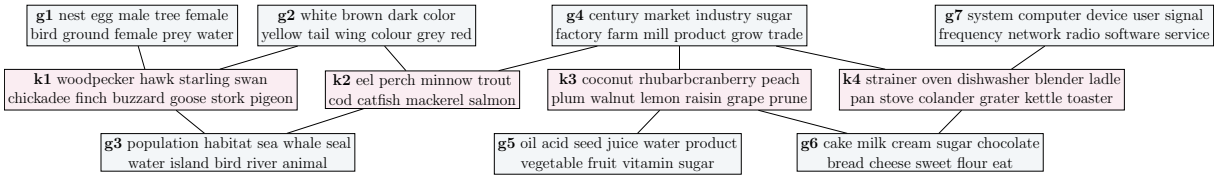


Figure 2: Examples of categories (red) and feature types (blue) inferred by BCF from the English Wikipedia. Connecting lines indicate that an association between the category and respective feature type was induced by the model.

	en	ar	zh	fr	ge
concepts	491	394	450	484	482
features	5,898	5,870	6,516	6,416	6,981
stimuli	418,755	86,908	147,386	258,499	233,175

Table 1: Datasets derived from Arabic (ar), Chinese (zh), English, (en), French (fr), and German (ge) Wikipedia.

are publicly available (Vinson & Vigliocco, 2008; Fountain & Lapata, 2010). Since the original studies were conducted in English, we collected translations of the target concepts and their categories into Arabic, Mandarin Chinese, French, and German from native speakers. We note that the final number of concepts featured in the language-specific corpora (Table 1) differs across languages for a number of reasons. First, some concepts get conflated as language differ in their polisemies (e.g., french *tongue* has two meanings (1) the organ *tongue* and (2) *language*) or conceptual granularity (e.g., everyday German does not distinguish between *mandarins* and *tangerins*). Second, some concepts were culturally less prevalent in some languages (e.g., *bagpipes* in Arabic) and hence not covered sufficiently in the resprective Wikipedia so that no input stimuli could be retrieved. Finally, some English concepts are expressed as multiple tokens, most prominently in Arabic (e.g., *ambulance* → سيارة اسعاف / literally: *ambulance car*), and our data processing pipeline may have missed some of these occurrences.

For each target language we created a corpus of input stimuli from articles from its respective Wikipedia dump;[4] we tokenized, POS-tagged and lemmatized the articles, and removed stopwords. From this data set we derived a set of input stimuli as target concept mentions in sentence context. In order to obtain balanced data sets, we automatically filtered words of low importance to a concept from contexts, using the term-frequency-inverse-document-frequency (tf-idf) metric. After filtering, we only kept stimuli with $3 \leq n \leq 20$ context words and at most 1,000 stimuli per target concept. Table 1 summarizes the statistics of the resulting data sets.

Experiments

Our experiments are designed to answer two questions: (1) Do computational models of category learning induce meaningful categorizations from rich and noisy data and across languages? We answer this question by applying three text-based categorization models to five diverse languages. We assess category quality by evaluating induced categories against a human-created reference categorization; and collect human judgments of the feature quality from large crowds of native speakers. (2) Does the capacity to learn structured features jointly with categories lead to qualitatively better representations? We answer this question by comparing BCF against BayesCat (a Bayesian cognitive categorization model with unstructured features) and co-occ (which is based on word co-occurrence). While all models retain a comparable category quality, humans judged BCF representation to be more *coherent* and *relevant*, and we present a qualitative analysis to corroborate this.

Parameters BCF and BayesCat learn K=40 categories and G=50 feature types. For both models we ran the Gibbs sampler for 1,000 iterations, and reported the final most likely representation. The co-occurrence model induces K=40 categories and G=5 feature types for each category. Reported numbers are averages over 10 runs.

Experiment 1: Category Quality

We compare model-induced categories against human created reference categorizations. We report purity (the extent to which each learnt cluster corresponds to a single gold class), collocation (the extent to which all members of a gold class are in a single learnt cluster) as well as their harmonic mean (F1 measure). We include a random baseline for reference which randomly assigns concepts to categories (averaged over 10 runs). Table 2 displays the results. BCF categories resemble the gold standard most closely, however, the difference to BayesCat is small for most languages suggesting that our joint category-feature model does not necessarily lead to higher quality categories. The categories learnt by the co-occ model are of lower quality across the board. Performance drops for languages other than English which is likely due to smaller stimuli sets (see Table 1).

Qualitative analysis. A few interesting idiosyncrasies emerge from our cross-lingual experimental setup, and the ambiguities inherent in language. For example, the English concepts *tongue* and *bookcase* were translated into French words *langue* and *bibliothèque*, respectively. The French BCF model induced a category consisting of only these two concepts with highly associated feature types {story, author,

[4] http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

	en	ge	fr	zh	ar
BCF	0.552 / **0.432 / 0.484**	0.454 / **0.400 / 0.425**	**0.534 / 0.441 / 0.483**	**0.441 / 0.349 / 0.389**	**0.408 / 0.327 / 0.363**
BayesCat	0.551 / 0.429 / 0.482	**0.458** / 0.378 / 0.414	0.507 / 0.415 / 0.457	0.430 / 0.320 / 0.367	0.394 / 0.298 / 0.339
co-occ	0.550 / 0.394 / 0.459	0.338 / 0.387 / 0.361	0.459 / 0.365 / 0.407	0.367 / 0.327 / 0.345	0.261 / 0.308 / 0.283
random	0.193 / 0.135 / 0.159	0.194 / 0.134 / 0.158	0.197 / 0.134 / 0.160	0.208 / 0.135 / 0.164	0.214 / 0.125 / 0.158

Table 2: Quality of induced categories in terms of purity / collocation / their harmonic mean (F1 measure) compared against the human-created gold standard. Bold indicates the best performing model for each language.

`saltwater soft naked marine family`				(target: *clam*)
BCF	snail	**clam**	moth	hare
Co-occ	level	snail	otter	whale

Figure 3: Illustration of the concept prediction task. Top left: model input (`features`). Top right: target prediction (*concept*). Bottom: top four predictions by BCF and co-occ.

	en	ge	fr	zh	ar
BCF	**34** (7)	31 (7)	**31** (7)	37 (9)	49 (13)
BayesCat	31 (5)	**32** (7)	28 (4)	**38** (9)	**54** (14)
co-occ	21 (1)	11 (0.7)	15 (2)	16 (3)	11 (1)

Table 3: Precision (%) at rank 10 (rank 1) over 300 test stimuli for BCF, BayesCat, and the co-occurrence model (co-occ). The number of concepts per language differs, so absolute numbers are not comparable across columns.

publish, work, novel} and {meaning, language, Latin, German, form}. Although this category does not exist in the gold standard, it arguably is a justifiable inference. Another example concerns the concept *barrel*, which in the English BCF output adopts its military sense and is grouped together with concepts *cannon, bayonet, bomb* and features like {kill, fire, attack}. Its French translation *baril*, however, refers to vessels exclusively and is categorized with *stove, oven* and represented through features including {oil, production, ton, gas}. Studying the interplay of language artefacts with conceptual knowledge is an interesting avenue for future work.

Experiment 2: Feature Quality

In order to assess the quality of induced feature representations, we confront each model with previously unseen stimuli d, and remove the concept. The models then rank all possible concepts $\hat{c}$ based on the given context features $\mathbf{f}_d$. Concept scores are computed as $s(\hat{c}|\mathbf{f}_d)=\sum_g P(g|\hat{c})P(\mathbf{f}|g)$ (BCF), $s(\hat{c}|\mathbf{f}_d)=\sum_{f\in\mathbf{f}_d} N(\hat{c},f)$ (BayesCat), and $s(\hat{c}|\mathbf{f}_d)=\sum_k P(\hat{c}|k)P(\mathbf{f}_d|k)$ (co-occ).

Figure 3 shows an English example stimulus, together with model predictions from BCF and co-occ. Results are shown in Table 3 as precision (%) at ranks 10 and 1 over a corpus of 300 unseen test stimuli. We again observe similar performance of BCF and the feature structure unaware BayesCat, while the co-occurrence model performs worse. We conclude that the joint and structured objective does not lead to feature representations that are more predictive for individual concepts. We further note that BCF and BayesCat perform comparatively across languages, while the performance of co-occ degrades, suggesting that the Bayesian models leverage information more efficiently from smaller input corpora (see Table 1). Considering the scarcity of categorization studies across diverse and potentially low-resource languages, robustness to small data sets is imperative.

Experiment 3: Human Judgments of Feature Relevance and Coherence

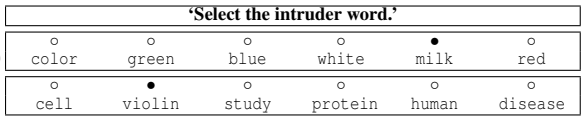
We are finally interested in how meaningful the induced feature representations are to humans, i.e., to native speakers of our five target languages. To this end, we elicited judgments of feature quality from native speakers of our five target languages using crowd sourcing. We adopted the topic intrusion experimental paradigm (Chang, Gerrish, Wang, Boyd-graber, & Blei, 2009) for assessing the induced features in two ways. Firstly, we examined whether the induced feature types are thematically *coherent*. Participants were presented feature types (as lists of words), which were augmented with a random 'intruder' feature (see Figure 4(a) left). The annotator was asked to identify the 'intruder feature'. If the feature types are internally coherent we expect annotators to complete this task with high accuracy. For each model (BCF, BayesCat and Co-occ), we evaluated all 50 induced feature types.

Secondly, we assessed the *relevance* of feature types to their associated categories. We presented participants with a category and five feature types (each as a list of words), one of which was randomly added and was *not* associated with the category in the model output (see Figure 4(b) left). If a model associates categories only with feature types that relevant, annotators will be able to identify the intruder with high accuracy. We evaluated 40 categories and their associated features for all three models.

For each of our target languages, we recruited annotators who self-identified as native speakers on crowd sourcing platforms Figure8 and Amazon Mechanical Turk.[5] We further filtered crowd workers by their location of residence, as well as by admitting only annotators with the highest platform-specific qualification level for all languages, and finally only admitted workers who correctly annotated a set of trial assignments. Each annotation task was completed by ten participants.

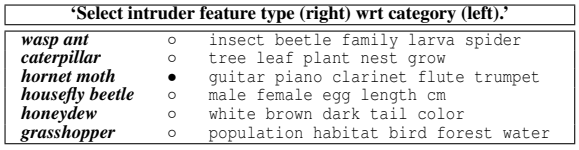
[5]All experiments were conducted on Figure8, except for the BayesCat evaluation for Chinese and French which was conducted on Amazon Mechanical Turk due to a decrease in the worker pool at the time of evaluation.

(a)



	en	ge	fr	zh	ar
BCF	**81** (0.71)	**76** (0.64)	**69** (0.53)	**57** (0.72)	**59** (0.44)
BayesCat	64 (0.53)	47 (0.34)	56 (0.41)	27 (0.23)	42 (0.42)
co-occ	27 (0.30)	22 (0.21)	22 (0.23)	23 (0.23)	24 (0.20)

(b)



	en	ge	fr	zh	ar
BCF	**75** (0.70)	**53** (0.36)	**56** (0.43)	**47** (0.42)	**39** (0.28)
BayesCat	61 (0.46)	37 (0.21)	47 (0.29)	26 (0.11)	**39** (0.22)
co-occ	30 (0.23)	34 (0.20)	28 (0.22)	39 (0.31)	28 (0.16)

Table 4: (a): Feature coherence study. Two example tasks (left) and human performance as accuracy (%) and Inter-annotator agreement (IAA; Fleiss Kappa) in brackets (right). (b): Feature relevance study. An example task (left; correct answer marked with filled circle), and human accuracy (%) and IAA in brackets (right).

Figure 4(a) displays the results for the *feature coherence* study (top) and Figure 4(b) shows the *feature relevance* results. Across both tasks BCF achieves best results by a large margin both in terms of accuracy and annotator agreement. We also observe a sharp drop in performance for languages other than English, and suspect that the smaller pool of crowd workers compounds the effect of a priori weaker representations (Tables 2,3). Overall, humans are able to detect intruder feature types more reliably in the context of BCF-induced representations, suggesting that modelling category acquisition jointly with structured feature induction leads to more *relevant* and internally *coherent* feature types as judged by native speakers. This result holds across all five languages, suggesting BCF as a valuable model for supporting exploratory, multilingual analyses of category-feature associations.

Qualitative analysis. Figures 2 and 4 qualitatively confirm that BCF learns meaningful features across languages,[6] which are overall coherent and relevant to their associated category. Figure 2 illustrates how feature types associate with different categories. For example g2 (`color`) is associated with both FISH (k2) and BIRDS (k1) whereas `bird habitat`-themed g1 is only associated with the latter; and FISH (k2), FRUIT (k3) APPLIANCE (k4) are all associated with `agriculture/industry`. Some interesting cultural differences emerge. For example German is the only language for which a `measurement` feature type is induced for VEGETABLES (Figure 4a; de, 4th from left), while for CLOTHING, a `fashion industry` feature emerges in French (Figure 4b; fr, 3rd from left). For the same category, a feature type pertaining to `colour` emerges for all five languages (bold margins). In addition, some learnt features are entirely language/culture-specific. For example, the 3rd feature type for VEGETABLES in Chinese (Figure 4a) includes the word 分 which refers to *the extent to which food is cooked* [7] and 烂 which is *the stage when food starts to fall apart after cooking (stewing)*. In addition, the feature types induced for the Chinese CLOTHING category include two words which both translate to the English word *wear*, but in Chinese are specific to *wearing small items* (e.g., jewelery; 戴), and *wearing clothes* (穿), respectively. Language-specific features are meaningful, and at the same time category-features associations across languages reflect culture-driven differences.

Discussion

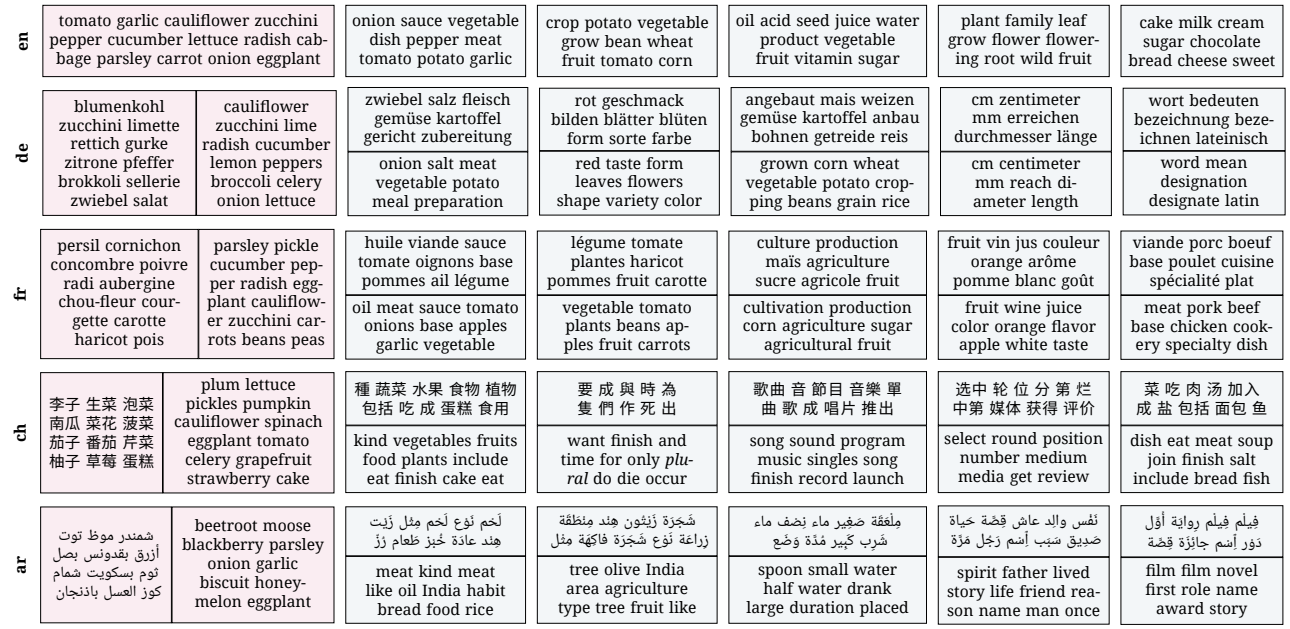
In this paper, we scaled computational cognitive studies of category acquisition to (a) five diverse languages; (b) larger concept and feature sets and noisier stimuli; and (c) the *joint* process of category and feature learning and presented a large-scale study of category and feature learning from natural language. Our results suggest that computational models of categorization learn meaningful categories and their features from rich and noisy data resembling the complexity of the world more closely than controlled laboratory settings. We also showed that BCF, a joint model of categories and structured features, induces the most *relevant* and *coherent* representations.

We conclude by discussing three avenues of future work. First, all our models formalize category acquisition as a general, language-independent process. They are unsupervised and neither utilize language-specific knowledge nor require custom parameter tuning. As such they pave the way for future investigations involving more languages, different genres and domains such as spoken language or fictional texts, or diachronic studies of the fine-grained change of conceptual representations drawing on historical data. Bayesian models are data efficient and expected to generalize to small data and low-resource languages.

Powerful language models (e.g., BERT; Devlin, Chang, Lee, and Toutanova (2019)) pre-trained on large text corpora in an unsupervised way, have been shown to encapsulate nuanced knowledge. While this work focused on cognitively motivated and interpretable models, a study of the extent to which powerful language models encapsulate conceptual knowledge is an interesting direction for future work.

[6] Model output of languages other than English was translated into English by native speakers.

[7] Roughly equivalent to the English *rare, medium, well-done*, but applicable to all kinds of food.



(a) Category CLOTHING

(b) Category VEGETABLES

Figure 4: Categories VEGETABLE (a) and CLOTING (b) (light red), and their five most highly associated feature types (light blue) for English (en), German (de), French (fr), Arabic (ar), and Chinese (ch). Model output of languages other than English was translated into English by native speakers.

Combining the strengths of both, e.g., by exposing BCF-like models to pre-trained word embeddings rather than raw text is also conceivable if target language resources permit.

Finally, even though our results corroborate prior work (Riordan & Jones, 2011) that the non-linguistic surrounding is to some extent encoded in language, Wikipedia is arguably a crude approximation of the environment which categories represent. We envision scalable testbeds which combine naturally occurring data from multiple modalities, for example combining text data with images or video. We demonstrated the potential of large naturalistic datasets for the development and testing of computational models, and are confident that computational cognitive models together with large naturally occurring data set will open up novel opportunities for investigating human cognition at scale.

References

Ahn, W.-K. (1998). Why are different features central for natural kinds and artifacts?: the role of causal status in determining feature centrality. *Cognition*, *69*, 135.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, *81*(3), 884–897.

Borovsky, A., & Elman, J. (2006). Language input and semantic categories: a relation between cognition and early word learning. *Journal of Child Leanguage*, *33*, 759-790.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Neurips.*

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories - feature predictability and information. *Psychological Bulletin*, *111*(2), 291–303.

De Melo, G., & Weikum, G. (2010). MENTA: Inducing multilingual taxonomies from Wikipedia. In *ACM CIKM.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL/HLT.*

Fountain, T., & Lapata, M. (2010). Meaning representation in natural language categorization. In *Cogsci.*

Fountain, T., & Lapata, M. (2011). Incremental models of natural language category acquisition. In *Cogsci.*

Frermann, L., & Lapata, M. (2015). A Bayesian Model for Joint Learning of Categories and their Features. In *NAACL/HLT.*

Frermann, L., & Lapata, M. (2016). Incremental bayesian category learning from natural language. *Cognitive Science*, *40*(6), 1333–1381.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual organization in vision: Behavioral and neural perspectives* (pp. 233–278).

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering Object Representations through Category Learning. *Cognition*, *78*, 27–43.

Gopnik, A., & Meltzoff, A. (1987). The Development of Categorization in the Second Year and Its Relation to Other Cognitive and Linguistic Developments. *Child Development*, *58*(6).

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Ji, L.-J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? examination of language effects in cross-cultural research on categorization. *Journal of Personality and Psychology*, *87*(1), 57–65.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3-36.

Malt, B. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85 – 148.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods*, *37*(4), 547–59.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3).

Medin, D. L., Unsworth, S. J., & Hirschfeld, L. (2007). Culture, categorization, and reasoning. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (chap. 25). Guilford Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700-708.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 328-350.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Cogsci.*

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When Categorization Flexibly Modifies the Perception of Faces in Rapid Visual Presentations. *Cognition*, *69*(3), 243 - 265.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 681–696.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1 - 25.

Spalding, T. L., & Ross, B. H. (2000). Concept learning and feature interpretation. *Memory & Cognition*, *28*, 439–451.

Vinson, D., & Vigliocco, G. (2008, February). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183–190.

Waxman, S. R., & Markov, D. B. (1998). Object properties and object kind: twenty-one-month-old infants' extensions of novel adjectives. *Child Development*, *69*, 1313-1329.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245–272.