black

white

The dozen or so boxes sat unopened in a garage for nearly 10 years. The archive of late cognitive anthropologist Robert E. MacLaury, they contained thousands of pages of handwritten survey results collected more than 35 years ago.

UC Irvine cognitive scientist Kimberly A. Jameson strongly believed their contents held key insights for researchers interested in learning how people form concepts to name and categorize color. She won a small grant from the UC Pacific Rim Research Program that allowed her to obtain the archive in 2011 and move it to UCI's Institute of Mathematical Behavioral Sciences, where she is an associate project scientist.

For Jameson, it is a treasure. "This is 30 years of raw data collected by Dr. MacLaury – what a shame for it not to be available to the public."

To understand why she felt this way, one has to look back to 1969 when seminal research on the theory of cross-cultural color categorization was published by anthropologist Brent Berlin and linguist Paul Kay, both of UC Berkeley. Their field study, which used missionaries to run color tests on indigenous societies they encountered in remote areas around the globe, collected color-naming data from speakers of 110 unwritten languages. Berlin and Kay believed that the order and way people identified color was universal, limited in part by human biology. Today the World Color Survey (WCS) is one of the most widely cited datasets in psychology.

MacLaury was a student and colleague of Berlin and Kay, and he went on to conduct his own color surveys. His archive includes the Mesoamerican Color Survey (MCS), a collection of irreplaceable observations of color-categorization behaviors gathered in 1978-81 from 900 monolingual speakers in over 116 indigenous languages.

MacLaury's survey is seen as a valuable extension of the WCS. "Rob's data was carefully collected since he either directed its collection, or personally collected it himself," says Jameson. "There is detailed documentation and all the techniques and instructions are standardized, which means there is a lot of regularity, and we can appreciate what's really going on comparatively."

Psychologists, linguists, anthropologists and cognitive scientists like Jameson study the evolution of color terms and how their meaning is understood and shared by societies over time. Researchers strive to capture how people think about and identify colors, aiming to precisely measure the behaviors and model them mathematically. The study of these data can thus potentially lend insight to a diverse range of human activities: organization and design of transportation systems, quality and risk in medical diagnoses, and physical and virtual design of retail markets and consumer goods, to name a few.

Jameson knew the MacLaury archive was valuable, and she felt a responsibility to share it.

It took three years to scan the archive's 23,000 pages. There were 142 different handwriting styles, and many surveys contained random notes and ethnographic

reports. But just preserving each page as a digital PDF was not enough. The next step is to make the surveys accessible online in an interactive format. Jameson envisions a resource in which researchers could easily query and manipulate the data to find linkages that could inspire original research and tools that would allow users to run their own simulation studies.

With nearly $1 million in support from the National Science Foundation, she assembled a team that included colleagues in cognitive sciences, mathematics, and ecology and evolutionary biology. She turned to Calit2 and its director, G.P. Li, to expand the team's technological expertise and for its experience in supporting multidisciplinary work.

"This project intrigued me. They wanted to apply Internet-based research methods to transform a paper archive," says Li, who wrote a letter of support in Jameson's NSF grant proposal. "It involves humanities, anthropology, information theory and computer sciences, a truly multidisciplinary project. It's perfect for Calit2, whose mission is to promote cross-disciplinary interaction with our own skill set in communications and information technology."

With the institute on board, the project gained Calit2 senior technology specialist Sergio Gago, as well as two undergraduate student research teams. Gago is working on a cloud-based collaborative framework, which runs a wiki engine *(Colcat.calit2.uci.edu)*, to house the digital archive. By modifying and enhancing the wiki framework, he is establishing the foundation on which to create a unique resource for users: a one-stop fully integrated platform that incorporates a host of tools, including maps, and search and simulation capabilities.

green

"It involves humanities, anthropology, information theory and computer sciences, a truly multidisciplinary project. It's perfect for Calit2, whose mission is to promote cross-disciplinary interaction with our own skill set in communications and information technology."

Meanwhile, student teams are tackling ways to efficiently transcribe the data. "This is our main challenge," says Gago.

They are writing optical character-recognition software to automatically convert handwriting into computer-addressable files. And they are designing the wiki to allow experts who are researching languages that have not yet been transcribed to download the files, and perform and enter the transcriptions themselves. In fact, each team member has transcribed one of the languages. However, the most promising approach to transcription has been crowdsourcing through Amazon's Mechanical Turk, an online job forum where people across the United States are able to log on and complete a human intelligence task, like transcribing a handwritten document.

Each document must be transcribed by several people to validate results. Prutha Deshpande, a cognitive science researcher working with Jameson on the project, explains that the team adapted a consensus-modeling approach to the Turkers' data.

Originally developed by UCI social scientists, the Cultural Consensus Theory is used by anthropologists to aggregate information and determine an agreed-upon truth. "We are expanding this model and applying it to perceptual tasks," says Deshpande.

Turns out, it's a really smart approach. "This model means we can reliably aggregate and get correct answers using crowdsourced data from small numbers of respondents," says Jameson. "Typically, one might need hundreds of inputs to obtain accurate 'majority-rule' answers. With this approach, we need fewer subjects' input to get the right answer with a high probability of certainty."

Jameson is happy with the progress. "We work well as a team, and some amazing things have happened."

It took the WCS about 23 years to convert and prepare for public distribution what had been recorded by the missionaries. The MacLaury archive contains many more languages, but is expected to be largely finished in three years.

"More importantly, we have developed a very robust and accurate modeling method by which to automate it and complete it quickly," says Jameson.

It is a technique that can be applied to other areas. She believes it's generalizable to any perceptually based crowdsource task that requires human participation, especially when machine learning is not adequate to provide reliable results. It may help with not

only other transcription projects, like deciphering soldiers' Civil War diaries, but also such tasks as looking at satellite images of the ocean in a search for downed or missing aircraft.

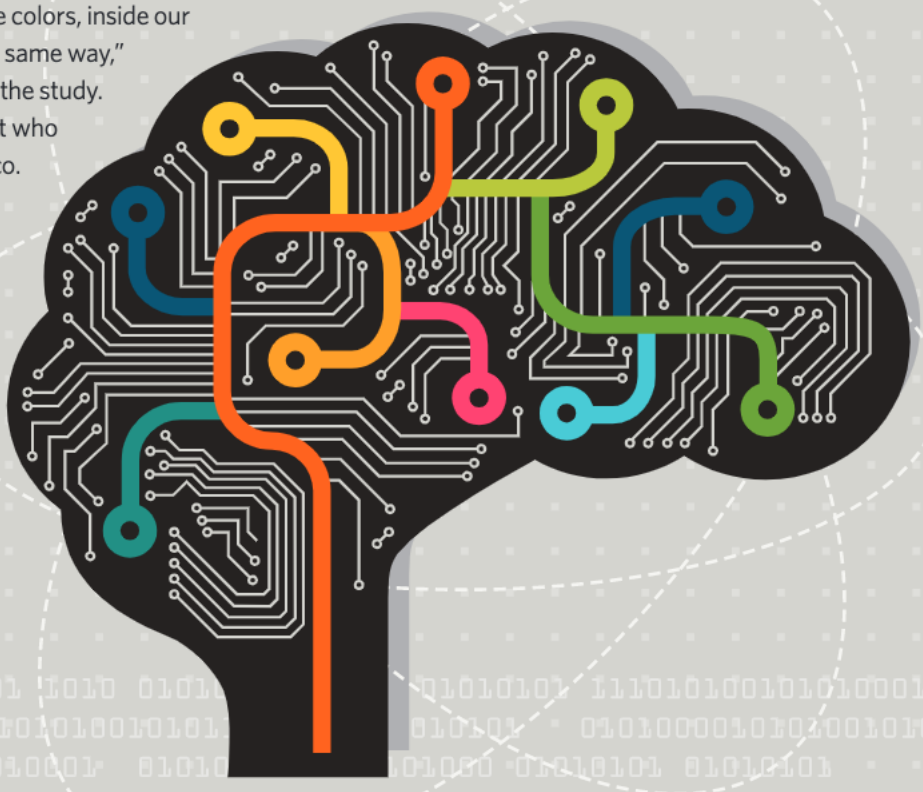
Jameson is also excited about her team's effort to develop simulated models of artificial populations and then study how they categorize color. She feeds her virtual societies information about what their environmental colors are, then speeds up time and sees through communication interactions how they would develop a categorization system.

"We want to understand how human concepts are formed and how they evolve, but science on this is constrained since researchers can't go back in time and observe what initially prompted the formation and sharing of a particular concept. Exploring concept evolution in artificial learners permits the testing and evaluation of factors that may have contributed to the shaping of our present human color concepts," she explains.

Two color researchers, Ohio State professors Delwin Lindsey (psychology) and Angela Brown (optometry), are looking forward to the database's public release. They have been studying individual differences in color naming within the same languages, using the WCS data, and they've found that color naming is not a matter of nature versus nurture, but a combination. Their results suggest that cultures create color names, but individuals from vastly different societies share the same understanding of colors in their mind.

"Though culture can influence how people name colors, inside our brains we're pretty much seeing the world in the same way," Lindsey said in an Ohio State news article about the study. "It doesn't matter if you're a native of Ivory Coast who speaks Abidji or a Mexican who speaks Zapoteco.

"We would like to test the conclusions of our analysis against another large dataset such as the MCS," said Lindsey. "This is a great service to the community of scientists who study the relationship between color and language. It will be an indispensable tool for those of us who investigate about how color-naming systems come into existence and how they change over time."



brown