Feature transmission within concept transmission

Stella Frank*[1] and Serge Belongie[1]

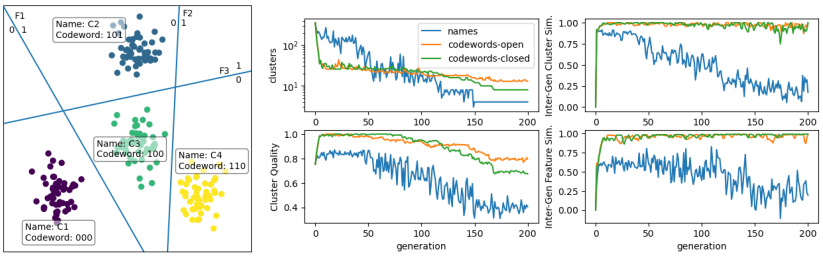
*Corresponding Author: stfr@diku.dk
[1]Pioneer Centre for Artificial Intelligence / DIKU, University of Copenhagen, Denmark

Conceptual categories can be described in terms of features: dogs (mostly) have four legs, bark, and shed, while clouds are grey or white and float in the sky. Concepts differ across cultures and languages (Everett, 2013; Majid, 2015), indicating a role for cultural transmission dynamics in their evolution (Contreras Kallens, Dale, & Smaldino, 2018; Carr, Smith, Culbertson, & Kirby, 2020). Since differences in conceptualisations are due to either using different feature boundaries or attending to different feature dimensions (e.g. weather-aware cultures may attend to cloud shape and color in a more fine-grained way), evolving the underlying feature space is integral to concept evolution. In this simulation-based study using Iterated Learning (IL) dynamics (Kirby, 2001), we show how the features underlying concept categories are co-evolved, *given a compositional signalling system* that surfaces the features as well as the concept extensions.

Previous work on concept evolution has focused on discovering concept extensions (Silvey, Kirby, & Smith, 2019; Carr, Smith, Cornish, & Kirby, 2017; Carr et al., 2020) but left the corresponding features implicit. We show that the features themselves, represented as boundaries in high dimensional space, can be also reliably transmitted as part of a compositional concept label.

Model In our framework (Fig. 1a), the world is represented as a high-dimensional *perceptual* feature space; objects are points in this space. *Semantic features* are linear decision boundaries (hyperplanes) in this space, distinguishing points on either side of the boundary. *Concepts* are the interior spaces delimited by the set of features. Borrowing from error correcting output codes (Dietterich & Bakiri, 1995), we represent a concept as a *codeword*, the bitstring representing the feature values corresponding to the concept. A concept also has a *name*, a categorical label. Codewords are by construction *compositional*, while names are *holistic*. While natural languages may not have codeword-like labels (cf. Kirby, Cornish, & Smith, 2008), concepts may be described in terms of their features.

In our IL setup, learner agents infer a semantic feature space, corresponding to a set of concepts, from (label, object) pairs. In the baseline *name* condition, the labels are holistic names, and the task is to learn, via a linear SVM, a hyperplane for



(a) Example Space (2D) (b) Simulation Results (64D)

Figure 1.: (a) Illustrative example of concept clusters in feature space. (b) Simulation results comparing compositional codewords to holistic names. Synthetic data has 26 well-separated clusters in 64 dimensions. Learners receive 100 labels, in the form of codewords or names, to learn (initally 20) features which they then generalise to 900 unlabeled items. Top left: number of clusters found by each learner (logscale); bottom left: similarity of found clusters to correct clusters, measured using VM (Rosenberg & Hirschberg, 2007); top right: cluster similarity between adjacent generations (learnability) using VM to evaluate the similarity of their labels on a test set; bottom right: feature similarity between adjacent generations, measured as average best-match cosine similarity of feature boundary vectors. Code available at `github.com/scfrank/ecoc_evolang24`.

each name that separates the items with that name from all other items (1-vs-rest). In the *codeword* condition, the labels are codewords composed of feature values. The agent given codewords learns a hyperplane for each feature (e.g., distinguishing items with 0 vs 1 in the nth codeword position), again using a linear SVM. To generate labels for new objects, for the next round of IL, agents use their feature space to determine the conceptual location of a new item (in other words, using the binary features to perform multiclass classification to generate a codeword). This can result in a novel codeword, if this combination of features did not appear in the agent's learning phase. In the 'open world' condition, these new concepts are passed as is to the next generation; in the 'closed world' condition, these novel codewords are mapped to the closest existing codeword using Hamming distance. In the *name* condition, items are always mapped to the closest existing named cluster. In the initial round, names and features are random. At each generation, uninformative features are removed, resulting in shorter codewords.

Results Our simulations (Fig. 1b) show that learning from names alone leads to agents with conceptual systems that are less stable and correspond less to the underlying world, compared to learning from codewords. Codewords also enable IL chains to preserve specificity, and have a natural way of creating new concepts (open-world setting) to counteract the transmission bottleneck.

Acknowledgements

This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1.

References

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, *41*(4), 892–923.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

Contreras Kallens, P. A., Dale, R., & Smaldino, P. E. (2018). Cultural evolution of categorization. *Cognitive Systems Research*, *52*, 765–774.

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, *2*, 263–286.

Everett, C. (2013). *Linguistic relativity: Evidence across languages and cognitive domains.* Walter de Gruyer.

Kirby, S. (2001). Spontaneous evolution of linguistic structure - an iterated learning model of the emergence of regularity and irregularity. *5*(2), 102–110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Majid, A. (2015). Comparing Lexicons Cross-linguistically. In J. R. Taylor (Ed.), *The Oxford Handbook of the Word.* Oxford University Press.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In J. Eisner (Ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 410–420). Prague, Czech Republic: Association for Computational Linguistics.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051.