Science
AAAS

Supplementary Materials for

From language development to language evolution: A unified view of human lexical creativity

Thomas Brochhagen *et al*.

Corresponding author: Thomas Brochhagen, thomas.brochhagen@upf.edu

Science **381**, 431 (2023)
DOI: 10.1126/science.ade7981

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

This document provides a full description of our materials and methods as well as supplementary information. "Data sets" describes the three main data sets our analyses build on. Section "Knowledge types" describes the four knowledge types that were used as potential predictors. This section also gives descriptive statistics on their relationship. "Model estimates and rankings" lays out model estimates and rankings. Section "Comparison of estimates across data subsets" reports on additional sensitivity analyses concerning different splits of the data. Section "Self- and cross-prediction" provides details on self- and cross-predictions, as well as on additional analyses to ensure, among others, that areal or phylogenetic biases did not influence our results. Section "Visual similarity from self-supervised model" reports on another set of robustness checks in which visual similarity is derived from a different model than the one reported on in the main text. Section "Visualizations" gives complementary visualizations that were backgrounded in the main text due to lack of space. Finally, Section "Repeated extensions" provides further information on the data points that appear in multiple data sets.

Data Availability

All the data and resources that our analyses build on are freely available. Our code repository (see below) provides details to obtain the data necessary to run the analyses. In their original form, the data is available from the following sources. The CLICS^3 data (*36*) is available at `https://clics.clld.org`. The overextension data from Ferreira-Pinto & Xu 2021 (*15*) is available at `https://github.com/r4ferrei/computational-theory-overextension`. The DatSemShift 3.0 data (*37*) is available at `https://datsemshift.ru/`. The English associativity data from *Small World of Words* (*29*), at `https://smallworldofwords.org/en/project`. The data from Visual Genome (*52*) is available at `https://visualgenome.org/`. The affectiveness data from Mohammad 2018 (*35*), at `https://saifmohammad.com/WebPages/nrc-vad.html`; and that from Warriner et al. 2013 (*34*), at `https://link.springer.com/article/10.3758/s13428-012-0314-x#SecESM1`. Taxonomical information from WordNet (*33*) was queried through NLTK (*53*), available at `https://www.nltk.org/howto/wordnet.html`. Glottolog data is available at `https://glottolog.org/meta/downloads`.

Code Availability

Data processing and analysis code used in this study is available at `https://osf.io/zkgu3/`

Materials and Methods

Data sets

Our analyses are based on three main data sets. The overextension data is from Ferreira-Pinto & Xu 2021 (*15*). It aggregates cases of overextension attested in the literature; e.g., from Rescorla

1980 (*6*) and Thomson & Chapman 1977 (*54*). Among others, these cases of overextension encompass recorded play sessions, diary records, and picture naming activities. The data is monolingual, from English speaking children.

The crosslinguistic colexification data is from the *Database of Cross-Linguistic Colexifications*, CLICS^3, (*36*). This large data set aggregates and normalizes typological word-meaning lists from other resources. For example, from the Intercontinental Dictionary Series (*55*). CLICS^3 relies on Glottolog (*56*) for language-level information, and provides information about meanings and their lexifications in over 3000 languages.

The crosslinguistic semantic change data is from a subset of the *Database of Semantic Shifts in the languages of the world 3.0*, DatSemShift, (*37*). DatSemShift also covers other non-creative or non-lexical processes such as borrowing or cognancy. We focus on the subset of shifts labeled as cases of "polysemy" since it corresponds to fossilized relics of lexical creativity.

We identify a language with its unique identifier from Glottolog (*56*), which also provides information about the macro-area in which it is spoken. CLICS^3 already comes enriched with this information. To add this information to DatSemShift we proceeded in two steps. First, all languages were automatically matched with their corresponding information from Glottolog if their names agreed in both resources (e.g., Spanish is called "Spanish" in both Glottolog and DatSemShift). Second, languages from DatSemShift with no match were manually matched (e.g., "Skolt Sami" in DatSemShift corresponds to "Skolt Saami" in Glottolog). After this process, we exclude data from languages that still lack an identifier since both the identifier itself and the geographic information it comes associated with are components of our models (see "Model estimates and rankings").

Following previous work on colexification (*10, 11*), data from English was excluded from both crosslinguistic data sets. We do so to minimize effects that may be due to having English as a language from which colexification or semantic change data comes from as well as as a source of information for the four knowledge type we consider (see "Knowledge types" for further detail and discussion; and "Self- and cross-prediction" and "Visual similarity from self-supervised model" for additional analyses testing for linguistic or geographic biases).

Knowledge types

We consider four knowledge types: associativity, vision, affectiveness and taxonomic closeness.

Associativity The associativity data is from the English Small World of Words project (*29*). It covers first, second, and third responses to 12,282 cues. Following De Deyne et al. (*29*), we use a decaying random walk-based measure to derive a measure of associativity from this data. This measure captures rich conceptual relations such as situation-based thematic relations (e.g., 'key'-'door'), and it has been shown to outperform alternative formulations in other semantic tasks (*29*) as well as for predicting crosslinguistic colexification patterns (*11*). It has also been shown to fare well at characterizing child overextension (*15*). The general idea is that the associative strength between cues is not reflected only by their shared responses but also through

more indirect relationships mediated by other cues and responses. To this end, the raw cue-response matrix is first normalized and re-weighted through pointwise mutual information. This matrix then serves as a basis to operationalize the associativity of two cues as a function of the number of responses that connect them in a network of cues, and their weight.

More precisely, let $\mathbf{P}$ be a PPMI-transformed and normalized cue-response matrix; $\mathbf{I}$ be the identity matrix; and $\alpha < 1$ a parameter that regulates to which extent the length of paths matters. An associativity matrix $\mathbf{G}$ is then obtained as follows:

$$\mathbf{G} = (\mathbf{I} - \alpha\mathbf{P})^{-1}. \tag{1}$$

The associativity of cues i and j is then

$$\text{associativity}(c_i, c_j) = \frac{\sum_k \mathbf{G}_{[r_k,c_i]}\ \mathbf{G}_{[r_k,c_j]}}{\sqrt{\sum_k (\mathbf{G}_{[r_k,c_i]})^2}\sqrt{\sum_k (\mathbf{G}_{[r_k,c_j]})^2}}, \tag{2}$$

with r_k indexing the column of response k. Following De Deyne et al., we use $\alpha = 0.75$. We refer to this work for explicit derivations and further discussion (*57*) and (*29*).

Visual similarity Visual resemblance is based on computationally-derived visual representations of meanings, following (*15*) and (*31*). For all English glosses of meanings found in the three data sets (see "Data sets") and not covered in Gualdoni et al. 2022 (*31*), all images with a matching name were retrieved from Visual Genome (*52*). Glosses with less than 30 matches were excluded to avoid sparse representations. These images were then processed by the state of the art language and vision model of Anderson et al. 2018 (*32*), yielding distributed (vectorial) representations of each image (see (*31*) and (*32*) for further detail). These images were then averaged to arrive at visual prototypes –or visual average representations– of meanings. For all the glosses covered in Gualdoni et al., we used their publicly available prototype representations, obtained through the same procedure. Following previous work (*15, 31*), visual similarity is then operationalized as the cosine similarity of such prototypes.

To ascertain that our results are not affected by the nature of the task that the model of Anderson et al. (*32*) engages in, through which it learns its visual representations, we additionally performed a sensitivity check. This check re-evaluates our main analyses employing a different source of visual information, taken from a model engaged in a task with no linguistic component. Our results are stable using either of the two models' visual representations. Details on this check and on the alternative model employed for this purpose are provided in "Visual similarity from self-supervised model", below.

Taxonomy Following Ferreira-Pinto & Xu 2021 (*15*), taxonomic similarity is based on the Wu-Palmer similarity of synsets in WordNet (*33*), with synsets being the basic representational unit of WordNet: a collection of synonyms that express the same meaning. This measure of similarity is the quotient of tree depth of the most specific meaning above both synsets times

two, and the sum of the depth of each of the synsets. That is, the Wu-Palmer similarity of synsets s_i and s_j is

$$\text{WuPalmer}(s_i, s_j) = 2\frac{\text{depth}(\text{lcs}(s_i, s_j))}{(\text{depth}(s_i) + \text{depth}(s_j))}, \quad (3)$$

with $\text{depth}(\cdot)$ being the depth of a synset in the taxonomy, and $\text{lcs}(\cdot, \cdot)$ the least common subsumer of two synsets. The least common subsumer is the most specific ancestor node to both. This yields a score bounded within 0 and 1 such that lower values correspond to meanings that are further apart in the taxonomy.

Affectiveness Following De Deyne et al. 2021 (*30*), affectiveness is operationalized as the cosine similarity of 9-dimensional vectors based on two resources. The first six dimensions are valence, arousal, and dominance ratings from male and female participants from Warriner et al.'s 2013 norms (*34*) . The remaining three dimensions correspond to rescaled valence, arousal, and dominance judgments from Mohammad 2018 (*35*). This is the sparsest knowledge type we consider, which may contribute to its lack of effect in characterizing crosslinguistic data (see Section "Model estimates and rankings").

Two important caveats apply to these four knowledge types. First, while they draw either from English-speaking subjects (associativity; affectiveness) or from English resources (Visual Genome; WordNet), we employ them to characterize crosslinguistic data beyond English in the case of colexification and semantic change. As discussed in the main text, this limitation is due to there not existing large scale resources for typologically diverse languages to cover all four knowledge types. The one exception is visual similarity. See "Visual similarity from self-supervised model" below for a scalable alternative that is less dependent on language. Second, as also discussed in the main text, all this data comes from adult language use whereas child overextension happens in early development. We hope that future research will produce language resources with a broader typological and developmental coverage to enable a more comprehensive analysis of the data.

Interestingly, while the anglo-centricity of some of these resources is a clear limitation, crosslinguistic colexification has been shown to be a good predictor of affectiveness (*18*). That is, colexification patterns were shown to provide a good basis to infer English affective norms. This finding suggest that, while there is certainly an important cultural component to this kind of data (*17*), it may be smaller than may be expected a priori –at least for the case of affectiveness.

Table S1 shows Pearson correlations between knowledge types for child overextension data; Table S2 does so for the data in CLICS^3; and, respectively, Table S3 does so for DatSemShift. As can be appreciated from the three tables, across data sets, knowledge types are not very correlated. In principle, this allows them all to contribute.

Model estimates and rankings

All models were diagnosed to rule out pathological estimates. We checked for sampling size

	Visual similarity	Associativity	Affectiveness	Taxonomy
Visual similarity	1.000	0.576	0.191	0.352
Associativity	0.576	1.000	0.239	0.473
Affectiveness	0.191	0.239	1.000	0.088
Taxonomy	0.352	0.473	0.088	1.000

Table S1: Pearson correlation between knowledge types in child overextension data.

	Visual similarity	Associativity	Affectiveness	Taxonomy
Visual similarity	1.000	0.280	0.001	0.068
Associativity	0.280	1.000	0.095	0.192
Affectiveness	0.001	0.095	1.000	0.067
Taxonomy	0.068	0.192	0.067	1.000

Table S2: Pearson correlation between knowledge types in colexification data (CLICS^3).

	Visual similarity	Associativity	Affectiveness	Taxonomy
Visual similarity	1.000	0.460	0.082	0.284
Associativity	0.460	1.000	0.142	0.359
Affectiveness	0.082	0.142	1.000	0.110
Taxonomy	0.284	0.359	0.110	1.000

Table S3: Pearson correlation between knowledge types in semantic change data (DatSemShift).

(> 0.001 effective samples per transition); that all parameters had a split $\hat{R} < 1.1$ (*58*); that they all had a Bayesian Fraction of Missing Information over 0.2; and that they lacked staturated trajectory lengths. All cross-validations had a shape parameter $k < 0.7$, suggesting reliable estimates (*38*).

All models are logistic regressions. The dependent variable is whether a word-referent pair participates in overextension; whether a meaning pair in a particular language participates in colexification (CLICS^3); or whether a meaning pair in a particular language participates in semantic change (DatSemShift). As predictors, models had one to four knowledge types (see "Knowledge types"). Models fit on CLICS^3 and DatSemShift additionally had population-level effects for language and macro-area (see "Data sets"). More precisely, in the case of colexification, for pair i and j in language l:

$$\text{colex}_{ijl} \sim \text{Binomial}(1, p_{ijl}), \tag{4}$$

$$\text{colex}_{ijl} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ colexify in } l, \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

The logistic regression with all four knowledge types then is as follows:

$$\text{logit}(p_{ijl}) = \beta_{0l} + \beta_1 \text{vis}(i,j) + \beta_2 \text{assoc}(i,j) + \beta_3 \text{tax}(i,j) + \beta_4 \text{affect}(i,j), \tag{6}$$

$$\beta_{0l} = \gamma_{00} + \beta_{01} \text{language}_l + \beta_{02} \text{macro area}_l, \tag{7}$$

with $\text{vis}(i,j)$ being the visual similarity of i and j; $\text{assoc}(i,j)$ their associativity; $\text{tax}(i,j)$ their taxonomic similarity; and $\text{affect}(i,j)$ their affectiveness. Population-level predictors *language* and *macro area* are nominal variables from Glottolog (see "Data sets"). The DatSemShift models have the same structure, only that they predict whether i and j participate in a shift in l. The overextension models, respectively, instead predict whether i and j participate in overextension; and do not have population-level equivalent to β_{0l} (Eq. 7) but instead have just an intercept β_0 that does not vary over languages or areas.

In the main analyses reported, we excluded data not covered by all knowledge types. Focusing on only data covered by all knowledge types could, however, introduce a bias. For instance, that all data points are visually representable could affect the contribution of affectiveness; argued to be particularly relevant for abstract meanings that lack a visual representation (*30*). To rule this out we conducted separate analyses using the largest subset of data available for each type on its own (Section "Comparison of estimates across data subsets"). These checks suggest that focusing on data points covered by all knowledge types has little impact on the results below: estimates derived from data sets covered by a single knowledge type agree with those from the more restrictive intersection covered by multiple knowledge types.

Child overextension models Table S4 ranks all models fit on the overextension data. It also indicates their predictive accuracy, using leave-one-out cross-validation (*38*). The first three

Model	$ELPD_{\Delta}$	ELPD	Accuracy
all predictors	0 (0)	-223 (14)	0.81
visual similarity, associativity, taxonomy	-1 (2)	-224 (14)	0.80
associativity, affectiveness, taxonomy	-8 (4)	-231 (14)	0.82
associativity, taxonomy	-11 (5)	-234 (13)	0.82
visual similarity, affectiveness, taxonomy	-12 (5)	-236 (13)	0.79
visual similarity, associativity, affectiveness	-14 (5)	-238 (13)	0.78
visual similarity, associativity	-16 (6)	-239 (13)	0.78
visual similarity, taxonomy	-16 (6)	-239 (13)	0.78
associativity, affectiveness	-24 (7)	-248 (13)	0.76
associativity	-27 (7)	-250 (13)	0.75
affectiveness, taxonomy	-43 (10)	-267 (12)	0.75
visual similarity, affectiveness	-44 (9)	-267 (12)	0.74
visual similarity	-49 (10)	-272 (11)	0.74
taxonomy	-54 (11)	-277 (11)	0.75
affectiveness	-110 (13)	-333 (6)	0.63

Table S4: Ranking of child overextension models in terms of expected log-predictive densities (ELPD, lower is better). $ELPD_{\Delta}$ is the estimated difference in ELPDs from the highest ranked model. Both ELPD and classification accuracy are estimated through leave-one-out cross-validation.

models rank similarly, but all uni- or bi-variate models are outranked. Table S5 gives the estimates from the best model.

Colexification models Table S6 ranks all models fit on the colexification data from CLICS[3]. It also indicates their predictive accuracy, using leave-one-out cross-validation (*38*). The first two models rank similarly. All other models are clearly outranked. Table S7 gives the estimates from the best model.

Semantic change models Table S8 ranks all models fit on the semantic change data (DatSemShift). It also indicates their predictive accuracy, using leave-one-out cross-validation (*38*). The first six models rank similarly, but all univariate models are outranked. Table S9 gives the estimates from the best model.

Comparison of estimates across data subsets

The following tables compare estimates derived from univariate models fit on the subset of data points covered by all four knowledge types (see "Knowledge types") against those from the larger subsets comprising all data points for which there was information concerning a

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.26	0.13	0.01	0.53
Associativity	1.04	0.22	0.63	1.49
Visual similarity	0.67	0.16	0.36	0.97
Affectiveness	0.29	0.14	0.03	0.56
Taxonomy	0.77	0.14	0.49	1.06

Table S5: Standardized estimates from the best overextension model.

Model	$ELPD_{\Delta}$	ELPD	Accuracy
visual similarity, associativity, taxonomy	0 (0)	-22730 (104)	0.75
all predictors	-2 (1)	-22732 (104)	0.75
visual similarity, associativity	-136 (16)	-22866 (104)	0.75
visual similarity, associativity, affectiveness	-136 (16)	-22866 (104)	0.75
associativity, taxonomy	-250 (22)	-22980 (104)	0.75
associativity, affectiveness, taxonomy	-251 (22)	-22981 (104)	0.75
associativity	-429 (29)	-23158 (104)	0.74
associativity, affectiveness	-430 (29)	-23159 (104)	0.74
visual similarity, affectiveness, taxonomy	-3308 (76)	-26038 (87)	0.68
visual similarity, taxonomy	-3350 (76)	-26079 (87)	0.68
visual similarity, affectiveness	-4129 (83)	-26858 (83)	0.67
visual similarity	-4217 (84)	-26947 (82)	0.67
affectiveness, taxonomy	-5787 (94)	-28517 (64)	0.61
taxonomy	-5868 (94)	-28598 (63)	0.61
affectiveness	-7985 (103)	-30714 (25)	0.55

Table S6: Ranking of colexification models ($CLICS^3$) in terms of expected log-predictive densities (ELPD, lower is better). $ELPD_{\Delta}$ is the estimated difference in ELPDs from the highest ranked model. Both ELPD and classification accuracy are estimated through leave-one-out cross-validation.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.60	0.16	0.29	0.96
Associativity	1.76	0.03	1.70	1.82
Visual similarity	0.32	0.01	0.30	0.35
Taxonomy	0.24	0.02	0.21	0.27

Table S7: Standardized estimates from the best colexification model ($CLICS^3$).

Model	$ELPD_{\Delta}$	ELPD	Accuracy
visual similarity, associativity, taxonomy	0 (0)	-1767 (34)	0.77
all predictors	-1 (1)	-1768 (34)	0.77
visual similarity, associativity	-4 (3)	-1771 (34)	0.78
visual similarity, associativity, affectiveness	-5 (3)	-1772 (34)	0.77
visual similarity, taxonomy	-9 (4)	-1776 (34)	0.66
visual similarity, affectiveness, taxonomy	-10 (4)	-1777 (34)	0.66
associativity	-15 (6)	-1782 (34)	0.77
visual similarity, affectiveness	-16 (6)	-1783 (34)	0.65
associativity, affectiveness, taxonomy	-359 (28)	-2126 (23)	0.77
associativity, taxonomy	-360 (29)	-2127 (23)	0.77
associativity, affectiveness	-391 (29)	-2159 (22)	0.77
visual similarity	-395 (29)	-2162 (22)	0.66
affectiveness,taxonomy	-512 (32)	-2279 (18)	0.60
taxonomy	-517 (32)	-2284 (18)	0.61
affectiveness	-613 (34)	-2380 (14)	0.59

Table S8: Ranking of semantic change models (DatSemShift) in terms of expected log-predictive densities (ELPD, lower is better). $ELPD_{\Delta}$ is the estimated difference in ELPDs from the highest ranked model. Both ELPD and classification accuracy are estimated through leave-one-out cross-validation.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.97	0.29	0.43	1.60
Associativity	2.23	0.11	2.01	2.45
Visual similarity	0.24	0.05	0.13	0.34
Taxonomy	0.16	0.05	0.06	0.26

Table S9: Standardized estimates from the best semantic change model (DatSemShift).

particular knowledge type. In other words, the former, more *restricted set*, comprises data points covered by all types: visual, affective, taxonomical and associative. The latter, *full set* comprises all data points for which we had visual information, in the case of a visual univariate model; affectiveness information, in the case of an affectiveness univariate model; and so on.

Table S10 contrasts models fits on these two data subsets for overextension. Table S11 does so for colexification; and Table S12 does so for semantic change. Overall, the difference between estimates across these two data splits is minimal. This suggests that focusing on the restricted set covered by all four knowledge types should not have a great impact on our analyses. Note however that slight differences are found in univariate models with taxonomy or affectiveness as predictors, in the case of crosslinguistic data (Table S11 and Table S12). Notwithstanding, taking the magnitude of these predictors' effects and the size of these differences into account, we conclude that the restricted set that covers all four knowledge types does not introduce much bias.

Self- and cross-prediction

As discussed in the main text, both self- and cross-prediction of the three best models are performed on data subsets that exclude pairs that appear in more than one data set. For instance, if the pair "red"-"blood" appears in both CLICS^3 and DatSemShift, it was removed when calculating classification accuracy in order to avoid carry-over due to witnessed pairs. Additionally, population-level effects of crosslinguistic data sets (see "Model estimates and rankings") were averaged out for cross-prediction on overextension. The reason is that neither crosslinguistic model has a population-level estimate for English, since English data was removed (see "Data sets").

We also conducted additional checks to ensure that self- and cross-prediction results were not due to phylogenetic or areal biases, either inherent in CLICS^3 and/or DatSemShift; or due to the overextension data being in English. To this end, the crosslinguistic models were re-fit, leaving out, one by one, each of the five major language families found within CLICS^3 (Northeast Caucasian, accounting for about 14% of the data; Indo-European for 14%; Austronesian for 8%; Sino-Tibetan for 7%; and Uralic for 5%) and DatSemShift (Indo-European, accounting for 38%; Afro-Asiatic for 8%; Northeast Caucasian for 8%; Austronesian for 8%; and Turkic for 7%). The same leave-one-out re-fitting process was conducted for all macro-areas: Eurasia; South America; Papunesia; Africa; North America; and Australia. Table S13 shows the cross-prediction results for colexification models; and Table S14 does so for semantic change models. These results are also visually presented in Figure 4B and 4C in the main text.

Relatedly, we checked whether leaving out all Indo-European data would have an effect on the cross-predictive capabilities of the three best models for each phenomenon that we report on in the main text (Fig. 2A-C; Table S5; Table S7; and Table S9). That is, this check asks how models that were exposed to all the data, including Indo-European, perform when predicting only the non Indo-European portion of the colexification and semantic change data. If our results concerning cross-prediction were driven by a bias due to English (for the overexten-

Child overextension	Estimate	Est.Error	Q2.5	Q97.5
Visual Similarity				
Restricted set				
Intercept	0.01	0.10	-0.20	0.22
Visual similarity	1.38	0.13	1.13	1.63
Full set				
Intercept	0.01	0.10	-0.19	0.22
Visual similarity	1.42	0.12	1.18	1.67
Associativity				
Restricted set				
Intercept	0.38	0.14	0.12	0.65
Associativity	2.03	0.20	1.66	2.44
Full set				
Intercept	0.48	0.12	0.26	0.71
Associativity	2.28	0.19	1.91	2.66
Affectiveness				
Restricted set				
Intercept	-0.03	0.09	-0.22	0.15
Affectiveness	0.67	0.12	0.45	0.91
Full set				
Intercept	-0.03	0.08	-0.18	0.12
Affectiveness	0.60	0.10	0.42	0.79
Taxonomy				
Restricted set				
Intercept	0.01	0.10	-0.20	0.21
Taxonomy	1.28	0.12	1.05	1.52
Full set				
Intercept	0.01	0.08	-0.14	0.16
Taxonomy	1.35	0.09	1.18	1.53

Table S10: Estimates from univariate models fit on the full child overextension data subset that covers each knowledge type individually, compared to the restricted subset that covers all knowledge types.

Crosslinguistic colexification	Estimate	Est.Error	Q2.5	Q97.5
Visual Similarity				
Restricted set				
Intercept	0.17	0.13	-0.10	0.44
Visual similarity	0.96	0.012	0.93	0.98
Full set				
Intercept	0.11	0.12	-0.12	0.33
Visual similarity	0.96	0.01	0.94	0.99
Associativity				
Restricted set				
Intercept	0.59	0.15	0.30	0.92
Associativity	2.11	0.027	2.05	2.16
Full set				
Intercept	0.34	0.05	0.24	0.45
Associativity	2.10	0.01	2.07	2.13
Affectiveness				
Restricted set				
Intercept	0.14	0.16	-0.18	0.48
Affectiveness	0.22	0.01	0.2	0.24
Full set				
Intercept	0.01	0.08	-0.15	0.18
Affectiveness	0.34	0.01	0.33	0.35
Taxonomy				
Restricted set				
Intercept	0.20	0.16	-0.14	0.52
Taxonomy	0.71	0.01	0.69	0.74
Full set				
Intercept	0.02	0.06	-0.10	0.14
Taxonomy	0.59	0.01	0.58	0.60

Table S11: Estimates from univariate models fit on the full colexification data subset that covers each knowledge type individually, compared to the restricted subset that covers all knowledge types.

Semantic change	Estimate	Est.Error	Q2.5	Q97.5
Visual Similarity				
Restricted set				
Intercept	0.60	0.33	-0.06	1.28
Visual similarity	0.85	0.04	0.77	0.93
Full set				
Intercept	0.53	0.30	-0.15	1.15
Visual similarity	0.88	0.04	0.80	0.96
Associativity				
Restricted set				
Intercept	0.98	0.27	0.46	1.56
Associativity	2.47	0.11	2.26	2.68
Full set				
Intercept	0.51	0.17	0.16	0.83
Associativity	2.87	0.05	2.77	2.98
Affectiveness				
Restricted set				
Intercept	0.73	0.39	-0.05	1.52
Affectiveness	0.20	0.04	0.13	0.27
Full set				
Intercept	0.17	0.10	-0.02	0.40
Affectiveness	0.27	0.02	0.23	0.30
Taxonomy				
Restricted set				
Intercept	0.63	0.34	-0.04	1.32
Taxonomy	0.57	0.04	0.49	0.65
Full set				
Intercept	0.04	0.04	-0.04	0.14
Taxonomy	0.64	0.02	0.61	0.67

Table S12: Estimates from univariate models fit on the full semantic change data subset that covers each knowledge type individually, compared to the restricted subset that covers all knowledge types.

Colexification models		
	Accuracy: Overextension	Accuracy: Semantic change
Excluded macro-area		
Africa	0.81	0.73
Australia	0.80	0.73
Eurasia	0.81	0.73
North America	0.81	0.74
Papunesia	0.81	0.73
South America	0.80	0.74
Excluded language family (glottocode)		
aust1307	0.81	0.74
indo1319	0.81	0.73
nakh1245	0.81	0.73
sino1245	0.81	0.74
ural1272	0.81	0.74

Table S13: Cross-prediction results for re-fitted colexification models when excluding particular macro-areas or language families.

sion model) or due to an Indo-European bias, more generally (for colexification and semantic change) we would expect the models' predictive capabilities to decrease. We find no such decrease. The best overextension model has a cross-predictive accuracy of 0.72 on colexification data without Indo-European (compared to 0.72 with Indo-European, see Fig. 2) and an accuracy of 0.73 on semantic change data without Indo-European (0.72 with, see Fig. 2). The best colexification model has an accuracy of 0.74 on semantic change data without Indo-European (0.74 with, see Fig. 2); and the best semantic change model scores 0.73 on colexification data with no Indo-European (0.74 with, see Fig. 2). These results are also visually presented in Figure 4A in the main text.

Taken together, these results suggest that cross-prediction results are stable and change little when removing information from particular regions or language families. The one exception is the exclusion of Indo-European (glottocode *indo1319*) when re-fitting the semantic change model, which reduces its cross-predictive accuracy for colexification by about 0.1 (Table S14). Interestingly, this model fares almost identically to the semantic change model with Indo-European when cross-predicting overextension data, with a drop of only about 0.01.

Visual similarity from self-supervised model

In all the analyses reported above, visual similarity estimates are based on representations from the language and vision model of Anderson et al. 2018 (*32*) (see "Knowledge types"). These representations are optimized on the task of object classification (in English) and attribute pre-

Semantic change models	Accuracy: Overextension	Accuracy: Colexification
Excluded macro-area		
Africa	0.81	0.74
Australia	0.81	0.74
Eurasia	0.78	0.75
North America	0.81	0.75
Papunesia	0.82	0.74
South America	0.81	0.75
Excluded language family (glottocode)		
afro1255	0.81	0.75
aust1307	0.81	0.74
indo1319	0.80	0.65
nakh1245	0.81	0.75
turk1311	0.81	0.75

Table S14: Cross-prediction results for re-fitted semantic change models when excluding particular macro-areas or language families.

diction, that is, the task of the model is to detect objects, predict their classes (i.e., their labels among a vocabulary of 1600 object names in VisualGenome), and predict their attributes (e.g., "red" or "striped"). This optimization happens in a supervised manner; that is, the model has access to human-annotated image-label pairs and learns to associate visual and linguistic information.

It is not clear that this operationalization is problematic for our purposes, because human perception has been shown to be influenced by language, as reviewed in (*59*). Still, to check that the language dependency that is introduced in the current model is not what is driving the results, we reproduce our results with less linguistically informed representations – those obtained via so-called self-supervised learning.

Self-supervised learning does not rely on any ground truth label (such as, for instance, image labels). Self-supervised models are well-known for the robustness and generalization capabilities of the representations they learn (*60–62*). These capabilities allow them to solve complex tasks related to different aspects of cognition (*63, 64*). In what concerns vision, self-supervised models learn image representations with the task of reconstructing the content of an image from distorted and cropped versions of it (*65, 66*). This training regime makes self-supervised vision models completely blind to linguistic information.

We use such a self-supervised model to test the robustness of our results. In particular, we do so to see whether our results are impacted by the choice between visual representations obtained through a linguistic task (supervised model, reported on in the main text) and those obtained from a non-linguistic task (self-supervised, reported on in the following).

We obtain a measure of visual similarity based on a self-supervised model by following the same procedure as for its supervised counterpart (see "Knowledge types"). To recapitulate, for all the English meaning glosses found in the three data sets, we retrieved images from Visual Genome with matching labels, sampling up to 500 images per gloss. Glosses with less than 30 matches were excluded to avoid sparse representations. We processed these images with the state of the art self-supervised model of Caron et al. 2021 (*65*). In this way, we obtained vector representations for each image. Since the model was trained in a self-supervised manner, it never had access to the names associated with the images. As before, we then grouped the images by gloss and averaged their representations to obtain a visual prototypes for each meaning. Visual similarity is then operationalized as the cosine similarity of visual prototypes. Note that the averaging step introduces a language dependency also in this case, because we are grouping the images to be averaged by gloss, that is, by their label in VisualGenome. However, this method is clearly less informed by language than the other one.

While they are derived in different ways, our two measures of visual similarity are correlated. Their Pearson correlation for the overextension data is 0.68; for colexification data it is 0.77; and for the semantic change data it is 0.76.

We re-fit the best models for each of the three phenomena that we report on in the main text (see Table S5; Table S7; and Table S9 for model coefficients, reproduced below for ease of comparison as Table S16; Table S18; and Table S20) keeping everything the same except for the visual similarity scores, now using the ones from the self-supervised model.

The results we obtain are very stable both in terms of the coefficients of each of the three models and in terms of self- and cross-predictive capabilities. This suggests that the added value of visual information for lexical creativity does not hinge on whether the visual embeddings are a product of the model performing a linguistic task (supervised) or not (self-supervised).

In more detail, the coefficients of the three models re-fit with self-supervised visual information are shown in Table S15 for overextension (cf. Table S16 with supervision); in Table S17 for colexification (cf. Table S18 with supervision); and in Table S19 for semantic change (cf. Table S20 with supervision). They are also visually summarized in Fig. S1A-C. For comparison, this figure is a direct counterpart of Fig. 2 in the main text, which shows results that involve the supervised visual representations. While, as could be expected, there are numerical differences with respect to the models that draw from supervised visual information (cf. Table S5; Table S7; and Table S9), they are very small. The main difference worth highlighting is that the colexification model relies slightly more on taxonomy than vision when the latter is derived from the self-supervised model, and vice versa for its supervised counterpart.

Importantly, as shown in Fig. S1D, self- and cross-prediction results barely change (cf. Fig. 2D in the main text). Self-prediction is stable for colexification and increases by 0.01 for overextension and semantic change. Cross-prediction on colexification data decreases by 0.01 and 0.02 for overextension and semantic change models, respectively. Cross-prediction on overextension data is the same for the semantic change model, irrespective of whether it draws from supervised or self-supervised visual information; and it decreases by 0.01 for the colexification model with self-supervised visual embeddings. Finally, cross-prediction on semantic

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.30	0.14	0.04	0.57
Associativity	1.25	0.22	0.84	1.69
Visual similarity	0.39	0.14	0.12	0.68
Affectiveness	0.31	0.14	0.04	0.59
Taxonomy	0.92	0.15	0.62	1.22

Table S15: Standardized estimates from overextension model with visual similarity derived in a self-supervised manner.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.26	0.13	0.01	0.53
Associativity	1.04	0.22	0.63	1.49
Visual similarity	0.67	0.16	0.36	0.97
Affectiveness	0.29	0.14	0.03	0.56
Taxonomy	0.77	0.14	0.49	1.06

Table S16: Standardized estimates from the overextension model with visual similarity derived in a supervised manner.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.58	0.16	0.26	0.91
Associativity	1.88	0.03	1.82	1.93
Visual similarity	0.20	0.01	0.17	0.22
Taxonomy	0.26	0.01	0.23	0.29

Table S17: Standardized estimates from colexification model with visual similarity derived in a self-supervised manner.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.60	0.16	0.29	0.96
Associativity	1.76	0.03	1.70	1.82
Visual similarity	0.32	0.01	0.30	0.35
Taxonomy	0.24	0.02	0.21	0.27

Table S18: Standardized estimates from colexification model with visual similarity derived in a supervised manner.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.98	0.31	0.42	1.61
Associativity	2.21	0.11	2.01	2.42
Visual similarity	0.46	0.05	0.37	0.56
Taxonomy	0.11	0.05	0.01	0.21

Table S19: Standardized estimates from semantic change model with visual similarity derived in a self-supervised manner.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.97	0.29	0.43	1.60
Associativity	2.23	0.11	2.01	2.45
Visual similarity	0.24	0.05	0.13	0.34
Taxonomy	0.16	0.05	0.06	0.26

Table S20: Standardized estimates from semantic change model with visual similarity derived in a supervised manner.

change data decreases by 0.01 for both overextension and semantic change models with self-supervised visual embeddings. In sum, all models still show a very high degree of carry-over from one phenomenon to another, indicating that our results are robust irrespective of whether visual information stems from a task with a linguistic component.

Visualizations

The following figures complement the ones shown in the main analyses. Figure S2 shows the univariate models that best explain different data points in each phenomenon. This figure visually underscores that that the four knowledge types largely account for complementary information. This is true across phenomena, and reflected by the fact that the clusters of data that are best accounted for by different knowledge types are relatively well delimited. That is, they have relatively little overlap between them. As illustrated by the four exemplary data points found in the corners of each panel, this does not mean that each data point is solely explained by a single knowledge type. For instance, while the overextension of 'bike' to mean "wheelchair" is best explained by the univariate visual similarity model (bottom-right corner of Fig. S2A), the associativity model also predicts the likelihood of overextension for this item to be high. In this way this figure also underscores that many cases overlap, being explained by a mixture of factors. This intuition is borne out by the fact that the best models for each phenomenon recruit information from multiple knowledge types, as shown by our main results and most clearly reflected in the model rankings (see "Model estimates and rankings").

Figure S3 is a counterpart of Figure 3 in the main text. The latter visually compares self-

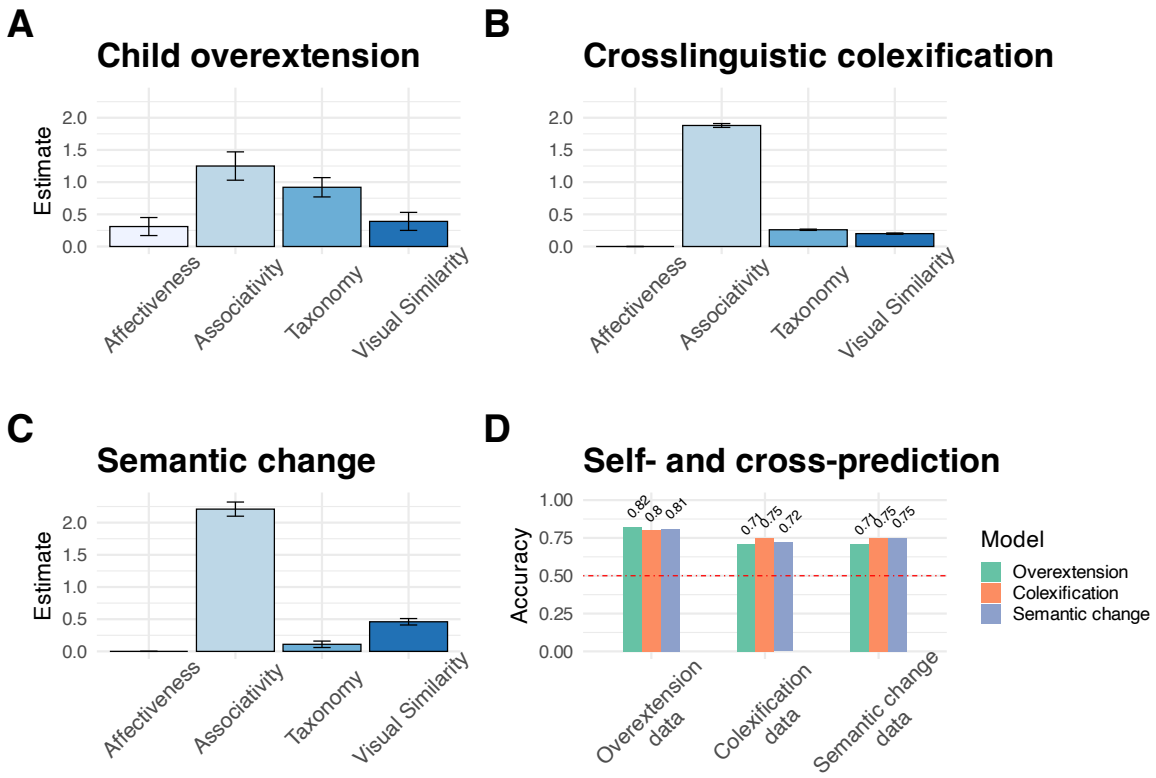

Figure S1: Summary of results with visual similarity derived in a self-supervised setting. A-C: Standardized estimates of effect of knowledge types from models of child overextension, colexification, and semantic change, respectively. As in the main text, while the models for evolutionary data (B-D) only use 3 predictors; we include a bar for affectiveness at 0 for illustration purposes. D: Accuracy of models when predicting new data. Self-predictions (e.g., colexification model's performance on colexification data) provide an upper-bound for cross-predictions. The random baseline of 0.5 (dashed line) provides a lower bound. Ceiling and bottom predictive accuracy are 1.0 and 0.

and cross-predictions across phenomena, showing predictions for attested cases of overextension, colexification, and semantic change. Figure S3 illustrates the same but for unattested cases. For attested cases model success implies assigning high probability to data points participating in lexical creativity. For unattested cases success implies the opposite: deeming lexical creativity unlikely. Accordingly, since predictive success lies in opposite directions for attested and unattested cases they are not visualized in a single figure. The general pattern remains the same in both cases: Model predictions generally pattern closely together, and agree both in success as well as in failure.

Repeated extensions

As described in the main text, some cases of lexical creativity are found in multiple data sets; and were excluded from the self- and cross-predictive analysis. When it comes to child overextension data, 57 cases appear in evolutionary data. This corresponds to about 22% of the data. In the case of the colexification data from CLICS^3, 3593 cases appear in overextension or DatSemShift. This corresponds to 16% of the data. In the case of the semantic change data from DatSemShift, 899 cases appear in overextension or CLICS^3. This corresponds to 50% of the data.

Table S21 lists all the child overextensions that are also found in either colexification or semantic change. Put differently, these cases correspond to crosslinguistically attested matches to instances of English child overextension. Table S22 gives the estimates from a model fit on this subset of data, using the same predictors as those of the two best models for evolutionary data: associativity, visual similarity, taxonomic similarity, and two population-level predictors for language and macroarea (see "Model estimates and rankings"; Table S6 and S8). On one hand, the model estimates are in line, both in terms of sign and ranking of effects, with the estimates derived from the larger data sets used in our main analyses (Table S7 and S9). On the other, their magnitudes –and the uncertainty about them– are very different. This is due to three combined factors. First, there is much less data to learn from. Second, as shown in Table S21, many meaning pairs appear in multiple languages. This means that many data points have the same prediction values, reducing the informativity of this data subset even further. Finally, there is also less information per language, adding further uncertainty about the effect of languages and macroareas. In sum, there is little information in this subset of the data to obtain reliable effect estimates. This result weakly signals that this data and its larger counterpart may be aligned; or at least not radically different (e.g., in suggesting sign reversals or differently ranked predictors).

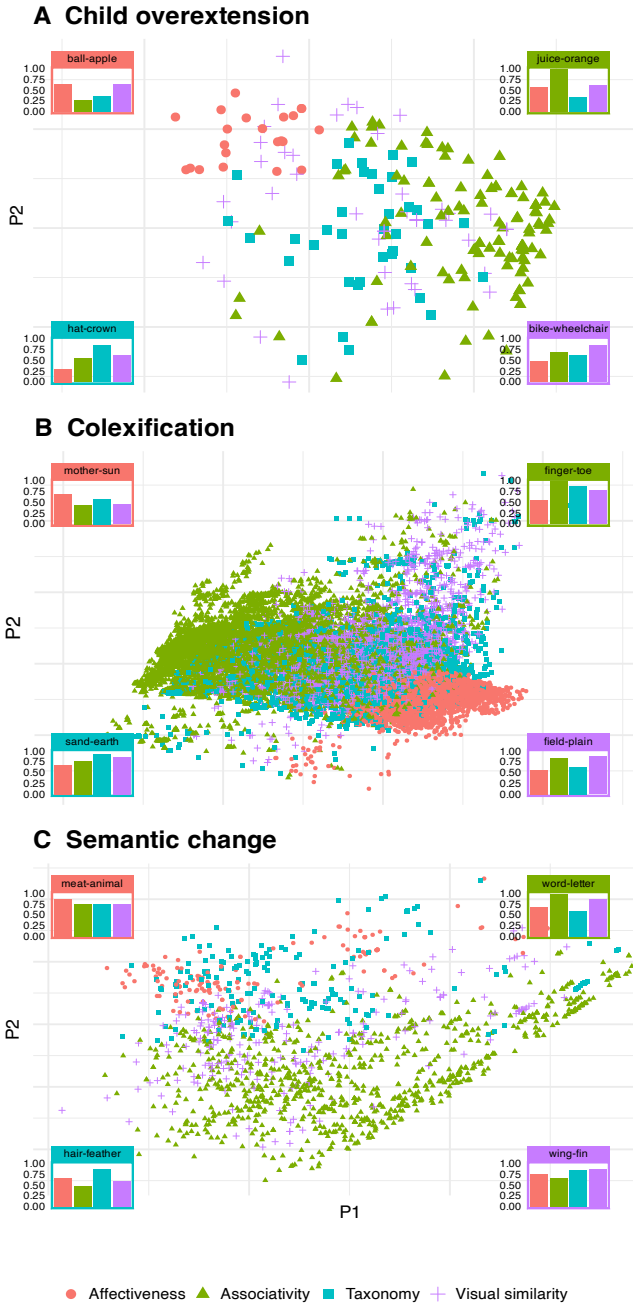

Figure S2: Illustration of multiple knowledge types playing a role in attested cases of (A) child overextension (B) colexification and (C) semantic change. Meaning pairs are projected onto two dimensions through principal component analysis. Data points correspond to correct predictions made by an univariate model using one of the four knowledge types. Color/shape indicate the type that predicts a given pair with the highest posterior probability. Bar plots in corners show posterior distributions of the four models for selected data points; box colors correspond to type with highest probability.

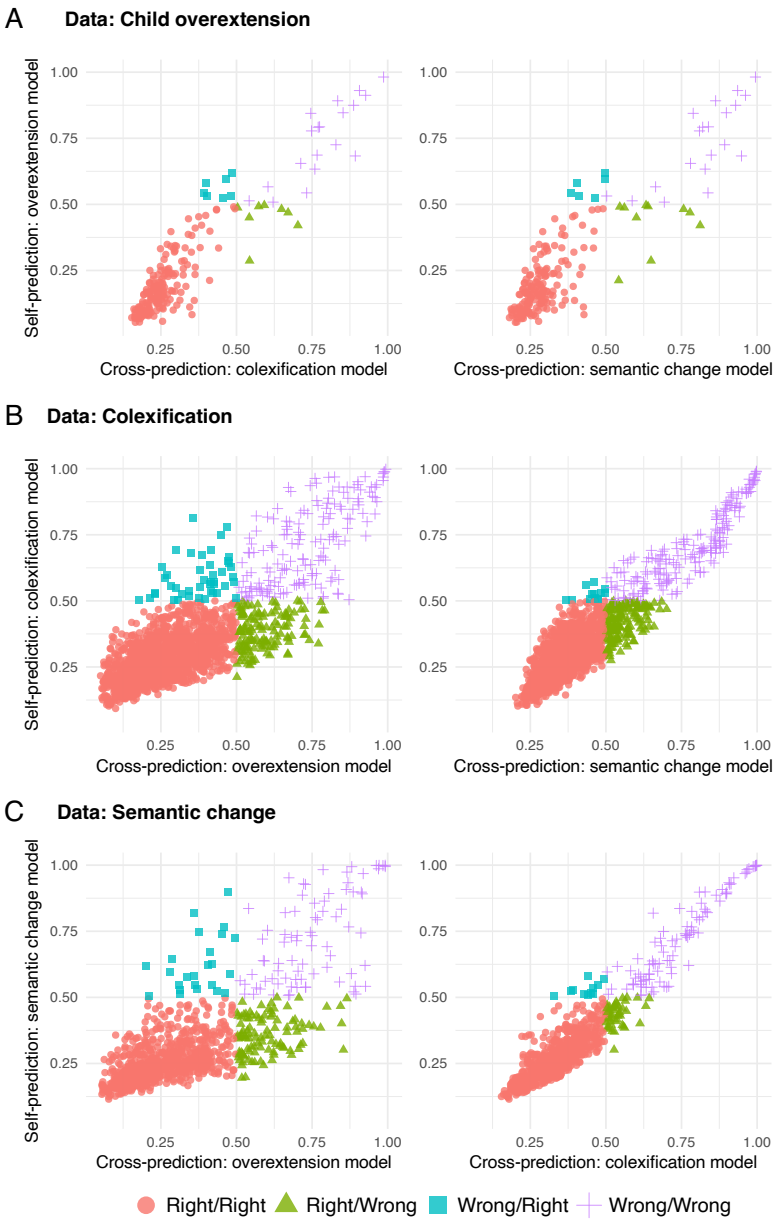

Figure S3: Comparisons of model self- and cross-predictions based on data from child overextension (A), colexification (B), and semantic change (C). Each panel compares self-predictions made by the best models of a phenomenon (y-axis) against cross-predictions made by the best models for another phenomenon (x-axis). Data points shown are unattested cases of overextension, colexification and semantic change. Colors and shapes separate predictions into classes: 'Right/Right' are correct predictions by both the self-predicting and cross-predicting models; 'Wrong/Wrong' are incorrect predictions by both; 'Right/Wrong' is a correct prediction from the self-predicting model but an incorrect one from the cross-predicting one, and conversely for 'Wrong/Right'. To make the plots legible, colexification data was randomly subsampled to 8%.

Glottocode	meaning1	meaning2	source
sart1249	bird	cow	CLICS
khoi1252	bird	cow	CLICS
guri1247	light	lamp	CLICS
guri1247	boot	shoe	CLICS
aika1237	moon	sun	CLICS
nucl1235	horse	donkey	CLICS
latv1249	baby	child	CLICS
lith1251	boot	shoe	CLICS
bulg1262	chicken	duck	CLICS
czec1258	boot	shoe	CLICS
poli1260	baby	child	CLICS
poli1260	boot	shoe	CLICS
bret1244	boot	shoe	CLICS
iris1253	boot	shoe	CLICS
goth1244	baby	child	CLICS
dutc1256	baby	child	CLICS
mode1248	baby	child	CLICS
sans1269	wagon	wheel	CLICS
sans1269	goose	duck	CLICS
sans1269	baby	child	CLICS
hind1269	baby	child	CLICS
vlax1238	baby	child	CLICS
beng1280	baby	child	CLICS
beng1280	goose	duck	CLICS
kash1277	moon	hair	CLICS
west2386	baby	child	CLICS
mara1378	light	lamp	CLICS
aves1237	wagon	wheel	CLICS
osse1243	baby	child	CLICS
west2369	baby	child	CLICS
stan1289	baby	child	CLICS
stan1290	baby	child	CLICS
roma1327	apple	fruit	CLICS
roma1327	boot	shoe	CLICS
stan1288	baby	child	CLICS
tokh1242	wagon	wheel	CLICS
bats1242	raft	boat	CLICS
lezg1247	boot	shoe	CLICS
cash1254	mother	father	CLICS

(continued)

Glottocode	meaning1	meaning2	source
tokh1243	wagon	wheel	CLICS
tokh1243	goose	duck	CLICS
hawa1245	light	lamp	CLICS
hawa1245	circle	ball	CLICS
hawa1245	baby	child	CLICS
maor1246	animal	dog	CLICS
maor1246	light	lamp	CLICS
rapa1244	light	lamp	CLICS
rapa1244	baby	child	CLICS
rotu1241	goose	duck	CLICS
rotu1241	elbow	knee	CLICS
rotu1241	boot	shoe	CLICS
tong1325	light	lamp	CLICS
chec1245	baby	child	CLICS
chec1245	elbow	knee	CLICS
kryt1240	baby	child	CLICS
khva1239	boot	shoe	CLICS
finn1318	baby	child	CLICS
udii1243	baby	child	CLICS
bezh1248	goose	duck	CLICS
bezh1248	baby	child	CLICS
khin1240	baby	child	CLICS
budu1248	boot	shoe	CLICS
avar1256	baby	child	CLICS
avar1256	boot	shoe	CLICS
lakk1252	baby	child	CLICS
lakk1252	boot	shoe	CLICS
darg1241	raft	boat	CLICS
darg1241	circle	ball	CLICS
darg1241	baby	child	CLICS
darg1241	boot	shoe	CLICS
kumy1244	baby	child	CLICS
kumy1244	boot	shoe	CLICS
taba1259	baby	child	CLICS
rutu1240	goose	duck	CLICS
rutu1240	boot	shoe	CLICS
iton1250	light	lamp	CLICS
mose1249	wagon	wheel	CLICS

(continued)

Glottocode	meaning1	meaning2	source
cavi1250	light	lamp	CLICS
tibe1272	baby	child	CLICS
jams1239	fish	dog	CLICS
waim1255	baby	child	CLICS
bara1380	moon	sun	CLICS
bari1297	elbow	knee	CLICS
bora1263	baby	child	CLICS
bora1263	moon	sun	CLICS
cara1272	baby	child	CLICS
chac1249	moon	sun	CLICS
chim1309	elbow	knee	CLICS
cube1242	moon	sun	CLICS
curr1243	baby	child	CLICS
desa1247	baby	child	CLICS
epen1239	eye	ball	CLICS
guah1255	moon	sun	CLICS
guam1248	baby	child	CLICS
hupd1244	baby	child	CLICS
hupd1244	bus	car	CLICS
hupd1244	moon	sun	CLICS
cacu1241	baby	child	CLICS
kore1283	baby	child	CLICS
muin1242	baby	child	CLICS
muin1242	moon	sun	CLICS
nuka1242	baby	child	CLICS
nuka1242	moon	sun	CLICS
ocai1244	baby	child	CLICS
play1240	moon	sun	CLICS
puin1248	baby	child	CLICS
resi1247	baby	child	CLICS
sion1247	baby	child	CLICS
sion1247	elbow	knee	CLICS
siri1274	baby	child	CLICS
tari1256	moon	sun	CLICS
tuyu1244	baby	child	CLICS
tuyu1244	boot	shoe	CLICS
mini1256	cat	dog	CLICS
mini1256	baby	child	CLICS

(continued)

Glottocode	meaning1	meaning2	source
muru1274	baby	child	CLICS
nupo1240	cat	dog	CLICS
woun1238	baby	child	CLICS
woun1238	moon	sun	CLICS
erzy1239	boot	shoe	CLICS
esto1258	baby	child	CLICS
khan1273	boot	shoe	CLICS
komi1268	baby	child	CLICS
mans1258	boot	shoe	CLICS
nene1249	boot	shoe	CLICS
selk1253	boot	shoe	CLICS
udmu1245	baby	child	CLICS
dido1241	boot	shoe	CLICS
dido1241	goose	duck	CLICS
enap1235	raft	boat	CLICS
macu1259	circle	ball	CLICS
macu1259	baby	child	CLICS
waiw1244	baby	child	CLICS
waiw1244	boot	shoe	CLICS
nege1244	baby	child	CLICS
jama1262	horse	donkey	CLICS
jama1262	boot	shoe	CLICS
high1278	baby	child	CLICS
karo1304	horse	dog	CLICS
seri1257	baby	child	CLICS
zuni1245	ball	bead	CLICS
nort2938	baby	child	CLICS
nucl1649	baby	child	CLICS
uppe1439	moon	sun	CLICS
uppe1439	elbow	knee	CLICS
sout2866	wagon	wheel	CLICS
yuwa1244	raft	boat	CLICS
yuwa1244	cat	dog	CLICS
pume1238	baby	child	CLICS
pume1238	boot	shoe	CLICS
cofa1242	baby	child	CLICS
waor1240	boot	shoe	CLICS
agua1253	circle	ball	CLICS

(continued)

Glottocode	meaning1	meaning2	source
agua1253	baby	child	CLICS
yagu1244	horse	donkey	CLICS
yagu1244	fork	spoon	CLICS
yagu1244	egg	door	CLICS
igna1246	baby	child	CLICS
trin1274	baby	child	CLICS
wapi1253	circle	ball	CLICS
wapi1253	baby	child	CLICS
wapi1253	boot	shoe	CLICS
waur1244	small	fruit	CLICS
cent2142	raft	boat	CLICS
cent2142	baby	child	CLICS
arao1248	light	lamp	CLICS
arao1248	baby	child	CLICS
esee1248	baby	child	CLICS
taca1256	light	lamp	CLICS
taca1256	circle	ball	CLICS
taca1256	baby	child	CLICS
pano1254	light	lamp	CLICS
pano1254	horse	donkey	CLICS
ship1254	circle	ball	CLICS
ship1254	baby	child	CLICS
yami1256	boot	shoe	CLICS
mund1330	moon	sun	CLICS
ache1246	boot	shoe	CLICS
east2555	ball	bead	CLICS
para1311	circle	ball	CLICS
para1311	baby	child	CLICS
siri1273	baby	child	CLICS
siri1273	horse	cow	CLICS
siri1273	cat	dog	CLICS
sana1298	raft	boat	CLICS
sana1298	baby	child	CLICS
sana1298	boot	shoe	CLICS
leng1262	baby	child	CLICS
moco1246	baby	child	CLICS
pila1245	circle	ball	CLICS
wich1264	baby	child	CLICS

(continued)

Glottocode	meaning1	meaning2	source
wich1264	circle	ball	CLICS
niva1238	butter	cheese	CLICS
niva1238	boot	shoe	CLICS
mapu1245	baby	child	CLICS
puel1244	light	lamp	CLICS
puel1244	boot	shoe	CLICS
qawa1238	light	lamp	CLICS
ghul1238	circle	ball	CLICS
ghul1238	lion	dog	CLICS
akhv1239	circle	ball	CLICS
akhv1239	baby	child	CLICS
akhv1239	boot	shoe	CLICS
bagv1239	baby	child	CLICS
botl1242	baby	child	CLICS
botl1242	fish	dog	CLICS
botl1242	boot	shoe	CLICS
kara1474	boot	shoe	CLICS
toki1238	baby	child	CLICS
toki1238	boot	shoe	CLICS
thai1261	horse	dog	CLICS
sout2746	baby	child	CLICS
nort2740	butter	cheese	CLICS
khun1259	butter	cheese	CLICS
guib1244	boot	shoe	CLICS
maon1241	baby	child	CLICS
hinu1240	boot	shoe	CLICS
emab1235	baby	child	CLICS
qaua1234	goose	duck	CLICS
tind1238	circle	ball	CLICS
tind1238	baby	child	CLICS
chir1284	baby	child	CLICS
cham1309	baby	child	CLICS
hunz1247	baby	child	CLICS
arch1244	baby	child	CLICS
kuba1248	boot	shoe	CLICS
kajt1238	baby	child	CLICS
east1436	baby	child	CLICS
east1436	goose	duck	CLICS

(continued)

Glottocode	meaning1	meaning2	source
east1436	boot	shoe	CLICS
tami1289	boot	shoe	CLICS
telu1262	elbow	knee	CLICS
noga1249	baby	child	CLICS
jude1256	baby	child	CLICS
manm1238	circle	ball	CLICS
manm1238	boot	shoe	CLICS
blan1242	circle	ball	CLICS
blan1242	boot	shoe	CLICS
huuu1240	baby	child	CLICS
huuu1240	horse	donkey	CLICS
boly1239	goose	duck	CLICS
maqu1238	elbow	knee	CLICS
elam1244	mother	father	CLICS
elam1244	mother	child	CLICS
tehu1242	baby	child	CLICS
gude1246	cow	father	CLICS
mafa1239	mother	father	CLICS
mamb1306	mother	father	CLICS
anga1288	mother	father	CLICS
clas1254	baby	child	CLICS
aleu1260	bird	duck	CLICS
chuk1273	bird	duck	CLICS
even1259	boot	shoe	CLICS
nana1257	bird	duck	CLICS
nucl1643	wheel	ring	CLICS
nucl1643	hook	key	CLICS
kett1243	bus	car	CLICS
kett1243	baby	child	CLICS
kild1236	baby	child	CLICS
sout2750	boot	shoe	CLICS
ainu1240	boot	shoe	CLICS
ainu1240	moon	sun	CLICS
zaiw1241	mother	father	CLICS
axiy1235	goat	dog	CLICS
kway1241	elbow	knee	CLICS
nden1248	horse	donkey	CLICS
nila1242	horse	donkey	CLICS

(continued)

Glottocode	meaning1	meaning2	source
zana1238	horse	donkey	CLICS
mach1266	leaf	banana	CLICS
mach1266	horse	donkey	CLICS
ngur1263	cat	dog	CLICS
ngur1263	goat	dog	CLICS
vwan1235	horse	donkey	CLICS
anja1238	elbow	knee	CLICS
araw1272	elbow	knee	CLICS
asas1240	elbow	knee	CLICS
bauz1241	moon	sun	CLICS
bepo1240	small	fruit	CLICS
buki1249	moon	sun	CLICS
erit1239	moon	sun	CLICS
fait1240	elbow	knee	CLICS
garu1246	moon	sun	CLICS
hula1239	elbow	knee	CLICS
kasu1251	moon	sun	CLICS
kesa1244	elbow	knee	CLICS
mala1495	elbow	knee	CLICS
male1291	elbow	knee	CLICS
kolo1267	elbow	knee	CLICS
muba1238	moon	sun	CLICS
musa1266	elbow	knee	CLICS
naka1265	elbow	knee	CLICS
nend1239	elbow	knee	CLICS
nend1239	small	fruit	CLICS
odoo1238	bird	dog	CLICS
pall1244	elbow	knee	CLICS
pamo1253	elbow	knee	CLICS
remp1241	mother	child	CLICS
rera1240	elbow	knee	CLICS
rumu1243	bird	dog	CLICS
samm1244	elbow	knee	CLICS
sile1255	elbow	knee	CLICS
wagi1249	baby	child	CLICS
wask1241	elbow	knee	CLICS
dump1243	elbow	knee	CLICS
dump1243	moon	sun	CLICS

(continued)

Glottocode	meaning1	meaning2	source
alek1238	elbow	knee	CLICS
awap1236	moon	sun	CLICS
biri1259	moon	sun	CLICS
dano1240	elbow	knee	CLICS
edop1238	moon	sun	CLICS
fayu1238	moon	sun	CLICS
gads1258	moon	sun	CLICS
kiri1256	moon	sun	CLICS
obok1239	moon	sun	CLICS
taus1252	moon	sun	CLICS
demi1242	moon	sun	CLICS
sika1261	flower	tree	CLICS
chew1245	fork	spoon	CLICS
gaww1239	baby	child	CLICS
imba1240	bus	car	CLICS
gali1262	fork	spoon	CLICS
cent2050	bus	car	CLICS
mana1288	goose	duck	CLICS
mezq1235	light	lamp	CLICS
sara1340	light	lamp	CLICS
yaqu1251	raft	boat	CLICS
tzot1259	bus	car	CLICS
bats1242	knee	elbow	DatSemShift
mika1239	peach	plum	DatSemShift
kich1262	deer	horse	DatSemShift
nucl1305	child	baby	DatSemShift
brah1256	child	baby	DatSemShift

Table S21: crosslinguistically attested colexifications or semantic shifts also found in the English child overextension.

Coefficient	Estimate	Est.Error	Q2.5	Q97.5
Intercept	1.31	0.71	0.11	2.95
Associativity	6.37	1.79	3.69	10.52
Visual similarity	2.10	0.63	1.03	3.52
Taxonomy	1.15	0.44	0.43	2.17

Table S22: Standardized estimates from model fit on crosslinguistically attested colexifications and semantic shifts that are also found in English child overextension.

References and Notes

1. M. Bréal, *Essai de Sémantique: Science des Significations* (Hachette, 1897).

2. S. Ullmann, Descriptive semantics and linguistic typology. *Word* **9**, 225–240 (1953). doi:10.1080/00437956.1953.11659471

3. C. Ramiro, M. Srinivasan, B. C. Malt, Y. Xu, Algorithms in the historical emergence of word senses. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2323–2328 (2018). doi:10.1073/pnas.1714730115 Medline

4. L. S. Vygotsky, *Language and Thought* (MIT Press, 1962).

5. E. V. Clark, Strategies for communicating. *Child Dev.* **49**, 953–959 (1978). doi:10.2307/1128734

6. L. A. Rescorla, Overextension in early language development. *J. Child Lang.* **7**, 321–335 (1980). doi:10.1017/S0305000900002658 Medline

7. E. V. Clark, *Cognitive Development and Acquisition of Language*, T. E. Moore, ed. (Academic Press, 1973), pp. 65–110.

8. A. François, *Studies in Language Companion Series* (JB, 2008), pp. 163–215.

9. A. Majid, J. S. Boster, M. Bowerman, The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition* **109**, 235–250 (2008). doi:10.1016/j.cognition.2008.08.009 Medline

10. Y. Xu, K. Duong, B. C. Malt, S. Jiang, M. Srinivasan, Conceptual relations predict colexification across languages. *Cognition* **201**, 104280 (2020). doi:10.1016/j.cognition.2020.104280 Medline

11. T. Brochhagen, G. Boleda, When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition* **226**, 105179 (2022). doi:10.1016/j.cognition.2022.105179 Medline

12. E. C. Traugott, R. B. Dasher, *Regularity in Semantic Change* (Cambridge Univ. Press, 2001).

13. E. Sweetser, *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure* (Cambridge Univ. Press, 1990).

14. D. P. Wilkins, *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, M. Durie, M. Ross, Eds. (Oxford Univ. Press, 1996), pp. 264–304.

15. R. Ferreira Pinto Jr., Y. Xu, A computational theory of child overextension. *Cognition* **206**, 104472 (2021). doi:10.1016/j.cognition.2020.104472 Medline

16. H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, T. Bhattacharya, On the universal structure of human lexical semantics. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1766–1771 (2016). doi:10.1073/pnas.1520752113 Medline

17. J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, K. A. Lindquist, Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019). doi:10.1126/science.aaw8160 Medline

18. A. Di Natale, M. Pellert, D. Garcia, Colexification networks encode affective meaning. *Affect. Sci.* **2**, 99–111 (2021). doi:10.1007/s42761-021-00033-1 Medline

19. G. Stern, *Meaning and Change of Meaning with Special Reference to the English Language* (Indiana Univ. Press, 1931).

20. L. Bloomfield, *Language* (Allen & Unwin, 1933).

21. D. Bickerton, *Language and Species* (Univ. of Chicago Press, 1990).

22. T. Givón, On the co-evolution of language, mind and brain. *Evol. Commun.* **2**, 45–116 (1998). doi:10.1075/eoc.2.1.04giv

23. D. Slobin, *Biology and Knowledge Revisited* (Routledge, 2004), pp. 273–304.

24. J. Culbertson, P. Smolensky, G. Legendre, Learning biases predict a word order universal. *Cognition* **122**, 306–329 (2012). doi:10.1016/j.cognition.2011.10.017 Medline

25. B. Beekhuizen, A. Fazly, S. Stevenson, Learning meaning without primitives: Typology predicts developmental patterns, *Proc. Ann. Conf. Cogn. Sci. Soc.* 170–175 (2014).

26. B. Beekhuizen, S. Stevenson, More than the eye can see: A computational model of color term acquisition and color discrimination. *Cogn. Sci.* **42**, 2699–2734 (2018). doi:10.1111/cogs.12665 Medline

27. S. Kirby, H. Cornish, K. Smith, Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10681–10686 (2008). doi:10.1073/pnas.0707835105 Medline

28. P. Monaghan, Age of acquisition predicts rate of lexical evolution. *Cognition* **133**, 530–534 (2014). doi:10.1016/j.cognition.2014.08.007 Medline

29. S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, G. Storms, The "Small World of Words" English word association norms for over 12,000 cue words. *Behav. Res. Methods* **51**, 987–1006 (2019). doi:10.3758/s13428-018-1115-7 Medline

30. S. De Deyne, D. J. Navarro, G. Collell, A. Perfors, Visual and affective multimodal models of word meaning in language and mind. *Cogn. Sci.* **45**, e12922 (2021). doi:10.1111/cogs.12922 Medline

31. E. Gualdoni, T. Brochhagen, A. Mädebach, G. Boleda, Woman or tennis player? Visual typicality and lexical frequency affect variation in object naming, in *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, J. Culbertson, A. Perfors, H. Rabagliati, V. Ramenzoni, Eds., 2022.

32. P. Anderson *et al.*, Bottom-up and top-down attention for image captioning and visual question answering, *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2018.

33. C. Fellbaum, *WordNet* (MIT Press, 2015).

34. A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013). doi:10.3758/s13428-012-0314-x Medline

35. S. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

36. C. Rzymski, T. Tresoldi, S. J. Greenhill, M.-S. Wu, N. E. Schweikhard, M. Koptjevskaja-Tamm, V. Gast, T. A. Bodt, A. Hantgan, G. A. Kaiping, S. Chang, Y. Lai, N. Morozova, H. Arjava, N. Hübler, E. Koile, S. Pepper, M. Proos, B. Van Epps, I. Blanco, C. Hundt, S. Monakhov, K. Pianykh, S. Ramesh, R. D. Gray, R. Forkel, J.-M. List, The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* **7**, 13 (2020). doi:10.1038/s41597-019-0341-x Medline

37. A. Zalizniak, *et al.*, Database of semantic shifts. *Moscow: Institute of Linguistics, Russian Academy of Sciences* (2016–2020); https://datsemshift.ru/.

38. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017). doi:10.1007/s11222-016-9696-4

39. S. De Deyne, G. Storms, Word associations: Network and semantic properties. *Behav. Res. Methods* **40**, 213–231 (2008). doi:10.3758/BRM.40.1.213 Medline

40. J. L. Bybee, D. I. Slobin, *Why Small Children Cannot Change Language on Their Own: Suggestions from the English Past Tense* (John Benjamins, 1982).

41. C. Kemp, T. Regier, Kinship categories across languages reflect general communicative principles. *Science* **336**, 1049–1054 (2012). doi:10.1126/science.1218811 Medline

42. C. Kemp, Y. Xu, T. Regier, Semantic typology and efficient communication. *Annu. Rev. Linguist.* **4**, 109–128 (2018). doi:10.1146/annurev-linguistics-011817-045406

43. Y. Xu, E. Liu, T. Regier, Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind (Camb.)* **4**, 57–70 (2020). doi:10.1162/opmi_a_00034 Medline

44. A. Karjus, R. A. Blythe, S. Kirby, T. Wang, K. Smith, Conceptual similarity and communicative need shape colexification: An experimental study. *Cogn. Sci.* **45**, e13035 (2021). doi:10.1111/cogs.13035 Medline

45. S. T. Piantadosi, H. Tily, E. Gibson, The communicative function of ambiguity in language. *Cognition* **122**, 280–291 (2012). doi:10.1016/j.cognition.2011.10.004 Medline

46. D. Gentner, M. Bowerman, *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, J. Guo, *et al.*, Eds. (Psychology Press, 2009), pp. 465–480.

47. J. M. Rodd, R. Berriman, M. Landau, T. Lee, C. Ho, M. G. Gaskell, M. H. Davis, Learning new meanings for old words: Effects of semantic relatedness. *Mem. Cognit.* **40**, 1095–1108 (2012). doi:10.3758/s13421-012-0209-1 Medline

48. M. Srinivasan, C. Berner, H. Rabagliati, Children use polysemy to structure new word meanings. *J. Exp. Psychol. Gen.* **148**, 926–942 (2019). doi:10.1037/xge0000454 Medline

49. Y. Xu, B. C. Malt, M. Srinivasan, Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognit. Psychol.* **96**, 41–53 (2017). doi:10.1016/j.cogpsych.2017.05.005 Medline

50. G. Zipf, *The Psycho-Biology of Language* (Houghton Mifflin, 1935).

51. T. Brochhagen, G. Boleda, E. Gualdoni, Y. Xu, From language development to language evolution: A unified view of human lexical creativity, OSF (2023); https://doi.org/10.17605/OSF.IO/ZKGU3.

52. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **123**, 32–73 (2017). doi:10.1007/s11263-016-0981-7

53. S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit (O'Reilly, 2009).

54. J. R. Thomson, R. S. Chapman, Who is 'Daddy' revisited: the status of two-year-olds' overextended words in use and comprehension. *J. Child. Lang.* **4**, 359–375 (1977).

55. The Intercontinental Dictionary Series, M. R. Key, B. Comrie, Eds. (Max Planck Institute for Evolutionary Anthropology, 2021).

56. Glottolog 4.8, Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian, Eds. (Max Planck Institute for Evolutionary Anthropology, 2020); http://glottolog.org.

57. S. De Deyne, D. J. Navarro, A. Perfors, G. Storms, Structure at every scale: A semantic network account of the similarities between unrelated concepts. *J. Exp. Psychol. Gen.* **145**, 1228–1254 (2016). doi:10.1037/xge0000192 Medline

58. A. Gelman, D. B. Rubin, Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.* **7**, 457 (1992).

59. G. Lupyan, R. Abdel Rahman, L. Boroditsky, A. Clark, Effects of Language on Visual Perception. *Trends Cogn. Sci.* **24**, 930–944 (2020). doi:10.1016/j.tics.2020.08.005 Medline

60. *Proceedings from NeurIPS 2019*, D. Hendrycks, M. Mazeika, S. Kadavath, D. Song, Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. Advances in Neural Information Processing Systems 32 (2019).

61. P. Goyal, et al., Self-supervised pretraining of visual features in the wild arXiv:2103.01988 [cs.CV] (2021).

62. X. Liu *et al*., Self-Supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* **35**, 857–876 (2023). doi:10.1109/TKDE.2021.3090866

63. T. Klein, M. Nabi, in *Proceedings of the Annual Meeting of ACL* (2020), pp. 7517–7523.

64. X. Zhu *et al*., in *Proceedings of the International Joint Conference on Artificial Intelligence* (2020).

65. M. Caron *et al*., in *Proceedings of IEEE International Conference on Computer Vision* (2021), pp. 9650–9660.

66. *Proceedings from NeurIPS 2020*, J.-B. Grill *et al*., Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems 33 (2020).