Colexifications for Bootstrapping Cross-lingual Datasets: The Case of Phonology, Concreteness, and Affectiveness

Yiyi Chen
Department of Computer Science
Aalborg University, Copenhagen
Denmark
yiyic@cs.aau.dk

Johannes Bjerva
Department of Computer Science
Aalborg University, Copenhagen
Denmark
jbjerva@cs.aau.dk

Abstract

Colexification refers to the linguistic phenomenon where a single lexical form is used to convey multiple meanings. By studying cross-lingual colexifications, researchers have gained valuable insights into fields such as psycholinguistics and cognitive sciences (Jackson et al., 2019; Xu et al., 2020; Karjus et al., 2021; Schapper and Koptjevskaja-Tamm, 2022; François, 2022). While several multilingual colexification datasets exist, there is untapped potential in using this information to bootstrap datasets across such semantic features. In this paper, we aim to demonstrate how colexifications can be leveraged to create such cross-lingual datasets. We showcase curation procedures which result in a dataset covering 142 languages across 21 language families across the world. The dataset includes ratings of concreteness and affectiveness, mapped with phonemes and phonological features. We further analyze the dataset along different dimensions to demonstrate potential of the proposed procedures in facilitating further interdisciplinary research in psychology, cognitive science, and multilingual natural language processing (NLP). Based on initial investigations, we observe that i) colexifications that are closer in concreteness/affectiveness are more likely to colexify; ii) certain initial/last phonemes are significantly correlated with concreteness/affectiveness intra language families, such as /k/ as the initial phoneme in both Turkic and Tai-Kadai correlated with concreteness, and /p/ in Dravidian and Sino-Tibetan correlated with Valence; iii) the type-to-token ratio (TTR) of phonemes are positively correlated with concreteness across several language families, while the length of phoneme segments are negatively correlated with concreteness; iv) certain phonological features are negatively correlated with concreteness across languages. The dataset is made public online for further research[1].

[1] https://github.com/siebeniris/ColexPhon

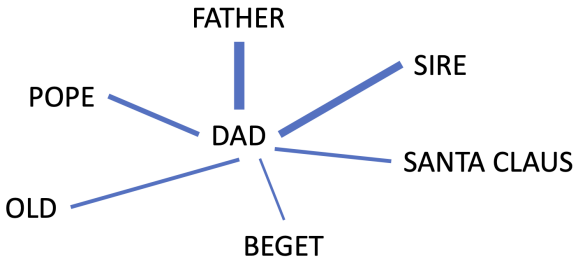

Figure 1: Colexification subgraph for DAD. The weight of the edges are proportional to the frequency of the colexification pattern in the dataset.

1 Introduction

Semantic typology studies cross-lingual semantic categorization (Evans et al., 2010). Within this area, the term “colexification" was first introduced and used by François (2008) and Haspelmath (2003) to create semantic maps. The study of colexifications focuses on cross-lingual colexification patterns, where the same lexical form is used in distinct languages to express multiple concepts. For instance, *mapu* in Mapudungun and *apakee* in Ignaciano both express the concepts EARTH and WORLD (Rzymski et al., 2020). Colexifications have been found to be pervasive across languages and cultures. The investigation of colexifications have led to interesting findings across different fields, such as linguistic typology (Schapper and Koptjevskaja-Tamm, 2022), psycholinguistics (Jackson et al., 2019), cognitive science (Gibson et al., 2019), but remain relatively unexplored in NLP (Harvill et al., 2022; Chen et al., 2023).

In recent years, with the increasing popularity of automatic methods and big data in linguistics, datasets such as Concepticon (List et al., 2022) and BabelNet (Navigli and Ponzetto, 2012) have been developed, affording large-scale cross-lingual semantic comparisons. The Database of Cross-lingual Colexifications (CLICS[3]) (Rzymski et al., 2020) was created based on the Concepticon con-

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 98–109

cepts, including 4,228 colexification patterns across 3,156 languages, to facilitate research in colexifications. Studies have also been shown to curate large-scale colexification networks from BabelNet, consisting of over 6 million synsets across 520 languages (Harvill et al., 2022; Chen et al., 2023).

While syntactic typology is relatively well-established in NLP (Malaviya et al., 2017; Bjerva and Augenstein, 2018a,b, 2021; Cotterell et al., 2019; Bjerva et al., 2019a,b,c, 2020; Stanczak et al., 2022; Östling and Kurfalı, 2023; Fekete and Bjerva, 2023), semantic typology has so far only been subject to limited research (Chen et al., 2023; Chen and Bjerva, 2023; Liu et al., 2023). As a relatively new topic in both semantic typology and NLP, colexifications covers a wide-range of languages and language families. In contrast, although the concepts of concreteness/abstractness and affectiveness (e.g., valence, dominance and arousal) have long been in the center stage of interdisciplinary research fields such as cognitive science, psychology, linguistics and neurophysiology (Warriner et al., 2013; Solovyev, 2021; Brysbaert et al., 2014), language coverage of such resources is severely limited, and curation prohibitively expensive.

The study of phonemes and phonological features have furthermore been essential to, e.g., address the problems of non-arbitrariness in languages and investigating universals of spoken languages (de Varda and Strapparava, 2022). Studies such as Gast and Koptjevskaja-Tamm (2022) demonstrate the genealogical stability (persistence) and susceptibility to change (diffusibility) via studying the patterns the phonemes/phonological forms and the colexifications across European languages. However, this study is limited to a small range of languages, and the investigated concepts are also restricted to 100-item Swadesh list (Swadesh, 1950). With the proposed procedures, a wider range of concepts and the phonological forms across language families are curated.

In this paper, we create a synset graph based on multilingual WordNet (Miller, 1995) data from BabelNet 5.0. We then develop a cross-lingual dataset that includes ratings of concreteness and affectiveness, as this approach yields more comprehensive data than using CLICS[3]. In addition, we meticulously select and organize phonemes and phonological features for the lexicons that represent the concepts. Our methodology for data creation is not limited to the constructed dataset, as it has potential for broader applications. We showcase the versatility of our approach through analysis across various dimensions, and make our dataset freely available.

2 Related Work

Colexifications The creation of semantic maps using cross-linguistic colexifications was initially formalized by François (2008). Semantic maps are graphical representations of the relationship between recurring expressions of meaning in a language (Haspelmath, 2003). This method is based on the idea that language-specific colexification patterns indicate the semantic proximity or relatedness between the meanings that are colexified (Hartmann et al., 2014). When analyzed cross-linguistically, colexification patterns can provide insights into various fields, such as cognitive principles recognition (Berlin and Kay, 1991; Schapper et al., 2016; Jackson et al., 2019; Gibson et al., 2019; Xu et al., 2020; Brochhagen and Boleda, 2022), diachronic semantic shifts in individual languages (Witkowski and Brown, 1985; Urban, 2011; Karjus et al., 2021; François, 2022), and language contact evolution (Heine and Kuteva, 2003; Koptjevskaja-Tamm and Liljegren, 2017; Schapper and Koptjevskaja-Tamm, 2022).

Jackson et al. (2019) conducted a study on cross-lingual colexifications related to emotions and found that different languages associate emotional concepts differently. For example, Persian speakers associate GRIEF closely with REGRET, while Dargwa speakers associate it with ANXIETY. The variations in cultural background and universal structure in emotion semantics provide interesting insights into the field of NLP. Bao et al. (2021) analyzed colexifications from various sources, including BabelNet, Open Multilingual WordNet, and CLICS[3], and demonstrated that there is no universal colexification pattern.

In the field of NLP, Harvill et al. (2022) constructed a synset graph from BabelNet to boost performance on lexical semantic similarity task. More recently, Chen et al. (2023) use colexifications to construct language embeddings and further model language similarities. Our goal is to utilize colexifications to construct cross-lingual datasets, including diverse ratings and phonological forms and features, to support further research, particularly in low-resource languages where norms and ratings are notably scarce.

Norms and Ratings A large number of words in high-resource languages have been assigned norms and ratings by researchers in psychology (Brysbaert et al., 2014; Warriner et al., 2013). Norms and ratings of words are essential components in psychology, linguistics, and recently being widely used in NLP. Norms refer to the typical frequency and context in which words are used in a particular language, while ratings represent subjective judgements of individuals on various dimensions such as concreteness, valence, arousal, and imageability. These norms and ratings can improve the performance on downstream tasks, such as sentiment analysis, emotion recognition, word sense disambiguation, and affective computing (Kwong, 2008; Tjuka et al., 2022; Strapparava and Mihalcea, 2007; Mohammad and Turney, 2010).

The study of concreteness and abstractness of concepts is interdisciplinary and spans across various fields, including linguistics, psychology, psycholinguistics, and neurophysiology (Solovyev, 2021). Concrete concepts are those that can be perceived by the senses, such as CAT and MOUNTAIN, while abstract concepts, like RELATIONSHIP and UNDERSTANDING, cannot be perceived by the senses. Brysbaert et al. (2014) conducted a study on concreteness ratings for 37,058 English words and 2,896 two-word expressions, involving over 4,000 participants, which has provided insights across various linguistic disciplines. The concreteness ratings are based on a scale of 1 (abstract) to 5 (concrete). These ratings have been used in conjunction with various tasks such as classification of metaphoricity (Haagsma and Bjerva, 2016) and animacy (Bjerva, 2014), as well as cultural studies (Berger and Packard, 2022).

Apart from concreteness, affective ratings are also essential for interdisciplinary research in psychology, linguistics and NLP. The affective norms for English words (ANEW) dataset, providing ratings of valence, arousal and dominance for English words, has been widely used in both psychology and NLP research (Bradley and Lang, 1999). Subsequently, the affective norms for French Words (FAN) and the affective norms for German words (ANGST) datasets, proving similar affective ratings for French and German words, respectively, have also been developed (Monnier and Syssau, 2014; Schmidtke et al., 2014). The Spanish version of ANEW is developed by Redondo et al. (2007). Extending the English ANEW, Warriner et al. (2013) covers nearly 14,000 English lemmas, providing ratings for valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus). For creating our dataset, we use the ratings from Warriner et al. (2013), see details in Section 3.

The data for linguistic norms and ratings is usually collected only for one language. For low-resource languages, such data is obviously lacking. Using our procedures, the norms and ratings can be bootstrapped for low-resource languages by sharing cross-lingual concepts through colexifications.

Phonemes and Phonological Features While direct phonetic comparison across languages is difficult, a common practice in comparing phonological characteristics across languages is to combine similar sounds into one multilingual phone set (Salesky et al., 2020). While more advanced methods for phonological typology do exist, e.g. Cotterell and Eisner (2017, 2018), a basic approach to phonology is found via the International Phonetic Alphabet (IPA), which classifies sounds based on general phonological properties. In this vein, WikiPron is created to serve as an open-source tool for mining phonemic pronunciation data from Wikitionary and still under continuous maintenance (Lee et al., 2020). To this date, it contains more than 1,8 million word/pronunciations across 543 languages.[2] The pronunciations are given in IPA, and segmented in a way that IPA diacritics can be properly recognized (Lee et al., 2020).

Demonstrating that phonological features outperform character-based models, PanPhon is created and used for various NER-related tasks (Mortensen et al., 2016). To date, PanPhon is a database relating over 5,000 IPA segments to 24 subsegmental articulatory features.[3] It has been used for various purposes, such as cross-modal and cross-lingual study of iconicity in languages (Zhu et al., 2021), and cross-linguistic phonosemantic correspondence using a deep-learning framework (de Varda and Strapparava, 2021).

In this paper, we build upon this work by diving into the relationship between phonological features, and the concreteness and affectiveness of sense lemmas across a wide set of languages. The paper is inspired by findings such that the sounds of words can influence their meaning and emotional

[2] https://github.com/CUNY-CL/wikipron
[3] https://github.com/dmort27/panphon

impact. For example, words with round vowel sounds are often associated with positive emotions, while harsher, more angular sounds can convey negative emotions (Ćwiek et al., 2022). This study aims to initiate the study on the intricate interplay between sound and affective/abstract meanings.

3 Dataset Curation

A *colexification pattern* refers to a case where two concepts are colexified, such as DAD-POPE shown in Figure 1. Specifically, a *colexification* is an instance of a *colexification pattern*, such as *far* in Danish, as shown in Table 1.

In order to leverage colexifications to create a cross-lingual dataset incorporating norms and ratings in psychology and other fields, we propose the following procedures for data curation and creation, as illustrated in Fig. 2.

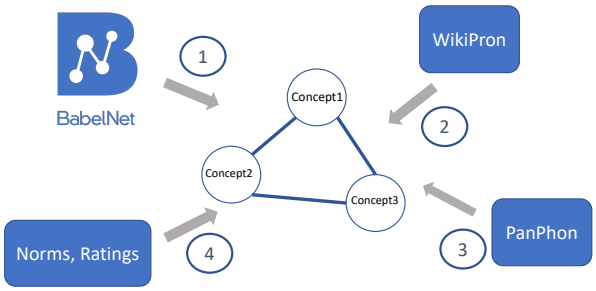

Figure 2: The Workflow of the Procedures for Creating the cross-lingual Dataset using Colexifications.

Building the Synset/Concept Graph In WordNet, a sense is a discrete representation of one aspect of the meaning of a word. For example, the lemma *bank* can either mean the sense FINANCIAL INSTITUTION or the sense SLOPING MOUND. The set of near-synonyms for a sense is called a **synset**, which is a primitive in WordNet (Jurafsky and Martin, 2023). Synsets are groups of words sharing the same concept. In order to construct of colexification networks, i) the WordDNet synsets are extracted from BabelNet; ii) for each synset, all the included word senses with their lemmas in the regarding language are elicitated; iii) finally, the sets of synsets sharing the same lemmas are extracted to represent a sysnet graph, with nodes being the synsets and the edges being the lemmas and their languages. The construction of a synset graph from BabelNet is first formalized in (Harvill et al., 2022), and adapted by (Chen et al., 2023) incorporating information of the languages and lemmas, see the Algorithm 1.

We adopt the algorithm presented in Chen et al. (2023) to construct a large-scale synset graph from WordNet synsets for our study. The difference in Chen et al. (2023) and Harvill et al. (2022) lies in the addition of G_s at line 3 and line 9, as shown in Algorithm 1. G_s affords the construction of colexification patterns and modeling language relations.

Algorithm 1 Construction of Colexification Graph: Given a set of languages L and corresponding vocabularies V, create graph edges between all colexified synset pairs (nodes), consisting of the set of tuples of lemmas and their language.

1: **function** CONSTRUCTGRAPH(L,V)
2: $CSP \leftarrow \{\}$ ▷ Colexified Synset Pairs
3: $G_s \leftarrow$ **graph**
4: **for** $l \in L$ **do**
5: **for** $x \in V_l$ **do**
6: **if** $|S_x| \geq 2$ **then**
7: **for** $\{s_1, s_2\} \in \binom{S_x}{2}$ **do**
8: $CSP \leftarrow CSP \cup \{s_i, s_j\}$
9: $G_s(s_1, s_2) \leftarrow \{x, l\}$
10: **end for**
11: **end if**
12: **end for**
13: **end for**
14: $G \leftarrow$ **graph**
15: **for** $s_1, s_2 \in CSP$ **do**
16: $G(s_1, s_2) \leftarrow 1$
17: **end for**
18: **return** G
19: **return** G_s
20: **end function**

A WordNet synset comprises a sense word, a Part-of-speech (POS) tag, and a sense number, e.g., dad#n#1. The sense numbers indicate the prevalence of the use of senses, with the most frequently used sense labeled 1. The frequency of use is determined by how often a sense is tagged in semantic concordance texts.[4] Our assumption is that the mean score of lexicon ratings, annotated by multiple humans across domains and languages, represents the ratings for the most prevalent sense. However, when it comes to cross-lingual synset-to-concept mapping, there may be variations in the sense annotations between languages. Suppose that in French the main sense KNOT is knot#n#4, which

[4] https://wordnet.princeton.edu/documentation/wndb5wn

refers to *a unit of speed*, while in English, the annotation for KNOT likely refers to *an actual knot that you tie*, which is the 1st sense for the synset. As a result, we cannot expect the same ratings of concreteness or affectiveness for these two different senses. Therefore, to map synsets to concepts, we always select the initial sense of the synsets..

Once filtered by the 1st sense of the synsets, as illustrated in Table 1, we derive concepts by extracting the sense word from each synset. The resulting concept graph comprises nodes representing the 1st senses of synsets and edges indicating the corresponding languages and sense lemmas.

Phonemes Extraction To facilitate analysis of phonetic characteristics cross-lingually in the context of colexifications and against ratings of concreteness and affectiveness, we extract phonemes from WikiPron, which to this date includes 1,882,240 word/pronunciation pairs in 543 languages.[5] To map the pronunciations to our data, we mapped their word/language code pairs to the pairs of sense lemma/language code extracted from BabelNet. As a result, there are 139,698 sense lemma/phonemes pairs across 142 languages, presented as in Table 1. In our dataset, the median size of the phonemes per language is 32.

Phonological Features Extraction Phonological features have been proposed as the foundation of spoken language universals. Despite variations in phones across languages, the set of phonological features remains constant. Phones can be constructed from a set of phonological features. In our study, we extract phonemes for sense lemmas and then further extract phonological (articulatory) features based on the subsegments using PanPhon. PanPhon generates 24 phonological features for each segment, such as syllabic, sonorant, consonantal, continuant, delayed release, lateral, nasal, strident, voice, spread glottis, constricted glottis, anterior, coronal, distributed, labial, high (vowel/consonant, not tone), low (vowel/consonant, not tone), back, round, elaric airstream mechanism (click), tense, long, hitone, hireg [6]. Each feature is assigned a value of '1', '-1', or '0', where '1' indicates a positive value of the feature, '-1' indicates a negative value of the feature, and '0' indicates that the feature is absent for that sound. For instance, a vowel cannot possess consonant features, so it is marked as '0'. We use PanPhon to convert each phone into a vector with length 24 in our dataset.

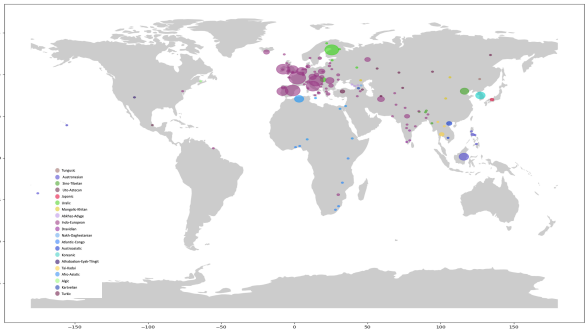

Figure 3: The map of language families of our data. The size of the points are proportional to the number of concepts in each language. Colors represent language families.

Incorporating Norms and Ratings Having built the concept graph from the synset graph by selecting the 1st senses of the synsets across languages, we map the concepts from databases containing norms and ratings to the concept graph. As shown in Table 1, the concept 1 DAD is mapped from concreteness/affectiveness rating lists to the synset 1 dad#n#1, while the concept 2 POPE is mapped to the synset 2 pope#n#1 by intersecting the datasets by the sense words. When each concept in the colexification pair has a rating, the distance of the concreteness/affectiveness can be calculated by computing the absolute distance of the two. When concept 1 has a (mean) concreteness of $conc_1$ and concept 2 has a (mean) concreteness of $conc_2$, then the $Conc.Dist$ is calculated as $|conc_1 - conc_2|$. Similar procedures are used for computing distance of valence ($V.Dist$), arousal ($A.Dist$) and dominance ($D.Dist$).

To conduct analysis of the correlations between phonemes/phonological features against the concreteness/affectiveness, the ratings for each phonemes are calculated as the average of the ratings of the included concepts, grouped by the phonemes and its language, respectively.

Undergoing these procedures, we create a dataset in 142 languages across 21 language families, including ratings in concreteness/affectivness, and phonemes for lemmas. The overall statistics of the data is shown in Table 2. The map for the data color coded by language families is presented in Fig. 3. As shown, the data is highly skewed towards Indo-European languages, and the data is quite scarce in Americas.

[5] https://github.com/CUNY-CL/wikipron
[6] https://github.com/dmort27/panphon

Sense Lemma	Language	Phonemes	Synset 1	Synset 2	Concept 1	Concept 2	Conc.Dist	V.Dist	A.Dist	D.Dist
پاپ	Persian	p ɑː p	dad#n#1	pope#n#1	DAD	POPE	0.42	1.96	0.16	1.88
بابا	Arabic	b ɑː b ɑː	dad#n#1	pope#n#1	DAD	POPE	0.42	1.96	0.16	1.88
папа	Russian	p a p ə	dad#n#1	pope#n#1	DAD	POPE	0.42	1.96	0.16	1.88
far	Danish	-	dad#n#1	sire#n#1	DAD	SIRE	-	0.74	0.05	0.57
pare	Castilian	p a r e	Santa_Claus#n#1	dad#n#1	SANTA CLAUS	DAD	0.17	-	-	-

Table 1: An example of the dataset. {CONC,V,D,A}.Dist represent the distance of the concreteness, valence, dominance and arousal of the pair of concepts for each lexicon. The value is unknown(-) if either of the concepts does not have a rating.

#Entries	Colex. Patterns	#Synset	#Lexicalization	#Phone/Lemma pairs	#Concept	#Concept w/ Aff.	#Concept w/ Conc.
186,6558	676,594	72,604	68,249	613,906	84,084	10,353	19,179

Table 2: Statistics of the Dataset.

4 Analysis and Results

4.1 Colexifications vs. Closeness in Concreteness/Affectivness

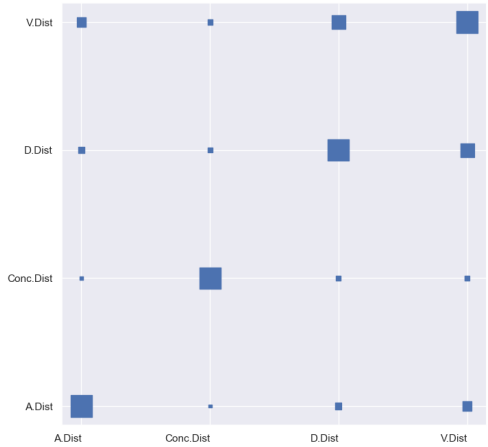

Figure 4: Correlation between Affectiveness- and Concreteness-Distances between the Colexified Concepts. The size of the squares represent correlation coefficients.

	Conc.Dist	V.Dist	A.Dist	D.Dist
#Colex.	-0.4716*	-0.4192*	-0.5798*	-0.5083*
Colex. Patterns	-0.4634*	-0.4115*	-0.581033*	-0.5065*
#Languages	-0.4727*	-0.4178*	-0.5798*	-0.5090*

Table 3: Correlation between #Colexifications and the Concreteness/Affectivness Distances between the Colexified Concepts, p-values are in the brackets. The sign * indicates the statistical significance of the correlation at 95% ($p < 0.0001$).

Previous studies show that abstract concepts are often understood by reference to more concrete concepts (Lakoff and Johnson, 2008), and words that first arise with concrete meanings often later gain an abstract one (Xu et al., 2017). Xu et al. (2020) leans on these findings to show that concepts more dissimilar in concreteness and affective valence are more likely to colexify. To test this, we calculate the correlation coefficients[7] between the number of colexifications and concreteness/affectiveness distances of the colexified concepts across languages. However, the results show the exact contrary to the previous theories and findings. As shown in Table 3, there is a statistically significant and relatively strong negative correlation between colexifications and the distance of concreteness, valence, arousal and dominance. This verifies that it is more likely for a pair of concepts to colexify when they are closer in concreteness and affectiveness. Our results about affectivness in colexifications is also corroborated by Di Natale et al. (2021).

Since both distances of conreteness and affectiveness are correlated with colexifications, it is intuitive to assume they might be correlated to each other. To test this, we calcuate the correlation coefficients between each dimension of concreteness and affectiveness. As shown in Fig. 4, the distances of valence and dominance are correlated with each other stronger than other pairs. And, concreteness distance is not significantly correlated with any dimension of affectiveness.

4.2 Phonemes vs. Concreteness/Affectiveness

Previous studies suggest that characteristics of the initial and the last phoneme have the most significant impact on the phonetic characteristics of the whole phone set (Pimentel et al., 2020). To test whether there are universals between the initial/last phoneme and the concreteness/affectiveness, we calculate the correlations between them per language family.

Since the whole results are too large to present,

[7] All the correlation analyses done in this study are using the SciPy implementation of Pearson correlation algorithm.

Lang. Family	#Lang.	# Sample	# Phonemes	Initial Phoneme	Last Phoneme
Turkic	7	2453	53	k (0.1148), t (0.1020)	-
Tai-Kadai	3	2701	20	k (-0.1122), n (0.1066)	-
Austroasiatic	2	3400	26	ʔ (0.1028)	-
Austronesian	7	21365	33	-	ŋ (0.1053)
Uralic	5	23352	37	ʊ (-0.1082)	i (0.1423), n (-0.1983), ɒ (0.1005)
Dravidian	3	339	22	p (0.2072)	ɭ (-0.2738)
Sino-Tibetan	5	7567	39	y (-0.1189)	[1] (-0.1428), [4] (0.1092), [5] (0.1066)
Afro-Asiatic	5	862	44	e (-0.1450)	o (-0.1107), r (-0.1582), β (-0.1074), χ (-0.1432)

Table 4: Correlation between the Initial/Last Phoneme and the Concreteness of Sense Lemma across Languages per Language Family. All the presented coefficients (in the brackets) are statistically significant and at least bigger than 0.1 or smaller than -0.1, corrected with Bonferroni correction ($p < 0.05/\#Lang.*$).

Features	Tai-Kadai (2822/3)	Austroasiatic (3555/2)	Indo-European (229661/75)	Uralic (26795/6)
syl	-0.1570*	-0.1870*	-0.1851*	-0.2716*
son	-0.1533*	-0.1698*	-0.1453*	-0.2783*
cons	-0.1734*	-0.2252*	-0.1284*	-0.2092*
cont	-0.1567*	-0.1768*	-0.1520*	-0.2692*
nas	-	-0.1038*	-0.1120*	-0.1718*
voi	-0.1524*	-0.1546*	-0.1726*	-0.2486*
sg	-	-0.1185*	-	-
ant	-0.1217*	-0.1407*	-0.1553*	-0.2670*
cor	-0.1574*	-0.1956*	-0.1215*	-0.2195*
distr	-	-	-	-0.1719*
lab	-	-	-	-0.1706*
lo	-	-	-	-0.1244*
hi	-0.1194*	-0.1678*	-0.1015*	-
lo	-0.1424*	-	-	-
back	-0.1009*	-0.1513*	-	-
tense	-0.1631*	-0.1175*	-0.1350*	-0.2675*

Table 5: Correlation between Phonological Features and the Concreteness of Sense Lemma per Language Family. All the presented coefficients are statistically significant and at least bigger than 0.1 or smaller than -0.1, corrected with Bonferroni correction ($p < 0.05/\#Lang.*$).

we report here only the results where the correlations are statistically significant, and the absolute value of which are bigger than 0.1. To prevent data from incorrectly appearing to be statistically significant, we correct the p-value with Bonferroni correction by dividing it with the number of the languages within the language family that is tested on. Only the results, that are statistically significant at 95% after applying Bonferroni correction, are reported.

We can observe that, as in Table 4, by correlating against the concreteness distance, the p as the initial phoneme and the last ɭ is significantly and stronger correlated within Dravidian languages, and a in Artificial languages as the first phoneme, compared to others. While across language families, k is correlated with concreteness.

Similarly, we test the correlations against the affectivness distance. Only the results with valence is reported, since the correlations of the phonemes against other affective ratings are not significant. As shown in Table 6, p as initials present correlations with affectiveness cross language families, i.e., Sino-Tibetan and Dravidian.

To represent the complexity of phonemes intra language families, we calculate the TTR as the ratio of unique phonemes and the length of all the phonemes for each lemma. Furthermore, the correlation between the TTR and the concreteness/arousal is computed, as shown in Table 4. And also the length of the phoneme segments are calculated for similar correlation test. Across all 8 language families, the segment length is statistically negatively correlated with the concreteness, but positively correlated with arousal. While, the correlations between TTR and the concreteness shows that the more concrete concept, the more diverse (complex) the phonemes are.

4.3 Phonological Features vs. Concreteness/Affectiveness

To test whether phonological features of the phonemes correlate with concreteness or affectiveness, for each phoneme/lemma pair, the phonological feature vectors are calculated and the values are aggregated by frequency of the present features. As indicated in Table 5, in the reported data, all the phonological features are negatively correlated with the concreteness. While the correlation coefficients in general are quite small, this hints at the possible existence of effects of these phonological features on concreteness. For instance, the *coronal obstruent (cor)* feature in all four language families is highly negatively correlated with concreteness, indicating that there is a general preference for such

Lang. Family	#Lang.	# Sample	# Phonemes	Initial Phoneme	Last Phoneme
Turkic	7	2453	53	c (-0.1178), a (-0.1284)	p (-0.1412), y (-0.1158)
Austroasiatic	2	3400	26	-	h (-0.1169)
Artificial Language	2	448	24	m (-0.2464)	-
Dravidian	3	339	22	p (0.1667), r (-0.2044)	ɭ (-0.2693)
Sino-Tibetan	5	7567	39	p (-0.1337), u (-0.1272), y (0.1010)	-
Afro-Asiatic	5	862	44	i (-0.1070), j (0.1065), z (-0.1058), g (-0.1268), ʔ (0.1091)	ɾ (0.1353), ʔ (-0.1588)

Table 6: Correlation between the Initial/Last Phoneme and the Valence of Sense Lemma across Languages per Language Family. All the presented coefficients (in the brackets) are statistically significant and at least bigger than 0.1 or smaller than -0.1, corrected with Bonferroni correction ($p < 0.05/\#Lang.$).

words to be abstract in meaning.

Lang. Family	#Lang.	# Sample	TTR	LEN
vs. Concreteness				
Turkic	8	2557	-	-0.1373*
Tai-Kadai	3	2701	0.1511*	-0.1834*
Austroasiatic	2	3398	0.1794*	-0.2715*
Uralic	6	23508	0.1876*	-0.2402*
Dravidian	3	339	-	-0.2585*
Indo-European	75	211371	-	-0.1697*
Sino-Tibetan	5	7567	0.1257*	-0.1184*
vs. Arousal				
Austroasiatic	2	3398	-	0.1157*
Mongolic-Khitan	3	66	-	0.3294*

Table 7: Correlation between TTR (Type-to-Token Ratio)/ Segment Length and the Concreteness of Sense Lemma per Language Family. All the presented coefficients (in the brackets) are statistically significant and at least bigger than 0.1 or smaller than -0.1, corrected with Bonferroni correction ($p < 0.05/\#Lang.$).

5 Conclusion and Future Work

In this study, we proposed a set of procedures to leverage colexifications to bootstrap cross-lingual datasets, incorporating human ratings of concreteness and affective meanings. The created dataset presents data in 142 languages across 21 language families and 5 language macro areas. However, the procedures can be applied beyond the datasets used in this paper.

Inspired by previous works, we test the correlations between i) the distance of concreteness/affectiveness and the number of colexifications; ii) the phonemes and concreteness/ affectiveness; and iii) the phonological features and the ratings. It is shown that i) colexifications closer in concreteness/effectiveness are more likely to colexify; ii) certian initial/last phonemes do present statistically significant correlations with the ratings across languages; and iii) there is a positive correlation between the phoneme diversity and concreteness; finally iv) certain phonological features are negatively correlated with the ratings. While it is difficult to draw any meaningful conclusions from this finding without a prior hypothesis, we hope that future work can use this dataset to make well-founded findings on the interactions between phonology, concreteness, and affectiveness.

We have showcased the soundness and validity of our approach to curate data from different domains and create a cross-lingual dataset mapping the information. The initial analyses and findings could inspire further applications in NLP and also other fields, such as psychology and psycholingusitics, which we will explore extensively for future work.

Nevertheless, the analyses conducted in this study are confined to individual correlation tests, which are inadequate for reaching definitive conlusions. For future work, we will employ multivariate modeling techniques utilizing affective/concrete ratings and the phonetic features to delve deeper into understanding the connections between human conceptualization and sounds across diverse languages and cultures.

Limitations

A limitation of this study is the fact that the concreteness ratings of Brysbaert et al. (2014) are curated solely from self-identified U.S. residents. And the affectiveness ratings of Warriner et al. (2013) are solely curated in English. As such, there is a risk of an anglocentric bias in the created dataset. Nonetheless, the goal of this study is to explore the potential of leveraging colexifications to bootstrap cross-lingual datasets in as many languages as possible, including a lot of low-resource languages.

Ethics Statement

Related to the limitations of this work, while this work increases research potential for low-resource languages, this comes with the main ethical risk of potential of propagating the anglocentric bias of some of the source datasets further.

Acknowledgements

This work is supported by the Carlsberg Foundation under a *Semper Ardens: Accelerate* career grant held by JB, entitled 'Multilingual Modelling for Resource-Poor Languages', grant code CF21-0454. We are furthermore grateful to the anonymous SIGMORPHON reviewers for pointing out issues that needed clarification in this work.

References

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference*, pages 1–7, University of South Africa (UNISA). Global Wordnet Association.

Jonah Berger and Grant Packard. 2022. Using natural language processing to understand people and culture. *American Psychologist*, 77(4):525.

Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution*. Univ of California Press.

Johannes Bjerva. 2014. Multi-class animacy classification with semantic features. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75.

Johannes Bjerva and Isabelle Augenstein. 2018a. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Johannes Bjerva and Isabelle Augenstein. 2018b. Tracking typological traits of uralic languages in distributed language representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86.

Johannes Bjerva and Isabelle Augenstein. 2021. Does typological blinding impede cross-lingual sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019b. Uncovering probabilistic implications in typological knowledge bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019c. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Johannes Bjerva, Elizabeth Salesky, Sabrina J Mielke, Aditi Chaudhary, Celano Giuseppe, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. Sigtyp 2020 shared task: Prediction of typological features. In *The Second Workshop on Computational Research in Linguistic Typology*, pages 1–11. Association for Computational Linguistics.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings.

Thomas Brochhagen and Gemma Boleda. 2022. When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, 226:105179.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Yiyi Chen, Russa Biswas, and Johannes Bjerva. 2023. Colex2Lang: Language embeddings from semantic

typology. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 673–684, Tórshavn, Faroe Islands. University of Tartu Library.

Yiyi Chen and Johannes Bjerva. 2023. Patterns of closeness and abstractness in colexifications: The case of indigineous languages in the americas. In *Third Workshop on NLP for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2018. A deep generative model of vowel formant typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 37–46, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390.

Andrea Gregor de Varda and Carlo Strapparava. 2021. A layered bridge from sound to meaning: Investigating cross-linguistic phonosemantic correspondences. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Andrea Gregor de Varda and Carlo Strapparava. 2022. A cross-modal and cross-lingual study of iconicity in language: Insights from deep learning. *Cognitive Science*, 46(6):e13147.

Anna Di Natale, Max Pellert, and David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science*, 2(2):99–111.

Nicholas Evans et al. 2010. Semantic typology. In *The Oxford handbook of linguistic typology*. Oxford University Press.

Marcell Richard Fekete and Johannes Bjerva. 2023. Gradual language model adaptation using fine-grained typology. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 153–158, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexandre François. 2008. Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, (106):163.

Alexandre François. 2022. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft*, 41(1):89–123.

Volker Gast and Maria Koptjevskaja-Tamm. 2022. Patterns of persistence and diffusibility in the european lexicon. *Linguistic Typology*, 26(2):403–438.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17.

Iren Hartmann, Martin Haspelmath, and Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 38(3):463–484.

John Harvill, Roxana Girju, and Mark Hasegawa-Johnson. 2022. Syn2Vec: Synset colexification graphs for lexical semantic similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5270, Seattle, United States. Association for Computational Linguistics.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language*, pages 217–248. Psychology Press.

Bernd Heine and Tania Kuteva. 2003. On contact-induced grammaticalization. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 27(3):529–572.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Dan Jurafsky and James H Martin. 2023. Speech and language processing (3rd (draft) ed.).

Andres Karjus, Richard A Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45(9):e13035.

Maria Koptjevskaja-Tamm and Henrik Liljegren. 2017. *Semantic Patterns from an Areal Perspective*, Cambridge Handbooks in Language and Linguistics, page 204–236. Cambridge University Press.

Oi Yee Kwong. 2008. A preliminary study on the impact of lexical concreteness on word sense disambiguation. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 235–244, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Johann Mattis List, Annika Tjuka, Christoph Rzymski, Simon Greenhill, and Robert Forkel. 2022. Clld concepticon 3.0.0 as cldf dataset.

Yihong Liu, Haotian Ye, Leonie Weissweiler, and Hinrich Schütze. 2023. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. *arXiv preprint arXiv:2305.12818*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Catherine Monnier and Arielle Syssau. 2014. Affective norms for french words (fan). *Behavior research methods*, 46(4):1128–1137.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12.

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online. Association for Computational Linguistics.

Antoinette Schapper and Maria Koptjevskaja-Tamm. 2022. Introduction to special issue on areal typology of lexico-semantics. *Linguistic Typology*, 26(2):199–209.

Antoinette Schapper, Lila San Roque, and Rachel Hendery. 2016. *12. Tree, firewood and fire in the languages of Sahul*, pages 355–422. De Gruyter Mouton, Berlin, Boston.

David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. 2014. Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior research methods*, 46:1108–1118.

Valery Solovyev. 2021. Concreteness/abstractness concept: State of the art. In *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 10-16, 2020, Moscow, Russia 9*, pages 275–283. Springer.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2022. Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods*, 54(2):864–884.

Matthias Urban. 2011. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1(1):3–47.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Stanley R. Witkowski and Cecil H. Brown. 1985. Climate, clothing, and body-part nomenclature. *Ethnology*, 24(3):197–214.

Yang Xu, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition*, 201:104280.

Yang Xu, Barbara C Malt, and Mahesh Srinivasan. 2017. Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive psychology*, 96:41–53.

Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1034–1041. IEEE.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations.