Pigments of our Imagination:
The Varied Semantics of Colour Term Colexifications

J.W. Brook

Vrije Universiteit Amsterdam

L_TAMATWS018: Comparative Linguistics

Prof. Dr. Antoinette Schapper

8th October, 2024

Abstract

A vast amount of scientific literature in the domain of lexical typology has revolved around a few subjects, seeming to have, in particular, a sort of fascination with colour terms, as well as terms of kinship and for body parts. This research attempts to explore much of the existing literature on colour semantics and colexification and carry out a quantitative cross-linguistic study into colour term colexification, with the aim of providing context to the rabbit-hole and analysing how various languages treat their basic colour terms. Using modern tools and linguistic databases, this study aims to draw stronger, source-based conclusions about how different language families treat their colour terms.

Keywords: colexification, colour semantics, lexical typology

Table of Contents

1 Introduction

In the Western world, we think of colour as one of the most fundamental aspects of human visual perception, serving as a crucial tool for distinguishing and categorising the world that we see (Wierzbicka, 2008). The assumption, it seems, is that our languages require standardised terms for basic colours so we can easily describe and communicate about these “innate” characteristics. The dominant *universalist* theory of colour semantics suggests that all languages treat colour in similar ways, and yet, not all languages even possess a term for “colour” or distinct terms used to describe individual colours, leaving the universality of colour terms open to question. This linguistic variability underscores that while colour perception is (generally) a universal human experience, the need for and use of specific colour terms can vary significantly between languages and cultures.

When we, in the West, think of certain colours, more than their innate properties spring to mind, instinctively picturing red as hot and blue as cold. This association is linguistic as well as visual, with both the word green and the colour green being inextricably linked to environmentalism for most English speakers (Lakoff & Johnson, 2008). Such associations highlight the complex interplay between language, culture, and perception in shaping our understanding of colours. Colour can often be thought of simply as a way to describe physical similarity between objects, but this description belies the complexity of colour perception.

From a physical perspective, colour perception should theoretically be consistent between observers, as colour can be described simply by the wavelength of a packet of electromagnetic radiation, with visible light’s wavelength falling between 380 and

700 nanometres. However, once these waves are processed as sensory input by the human eye, they are transformed by the brain into the rich, subjective experiences we associate with different colours. How these perceptions of colour are interpreted by different human brains is a complex and hotly-debated topic, but the term used to refer to these feelings of subjective, individual experiences of sensory perceptions in one's consciousness is *Qualia* (s.g. *quale*). Qualia are the ”what it's like” aspect of experiencing phenomena (e.g. the taste of chocolate or the perceived *blueness* of the sea) (Kanai & Tsuchiya, 2012). This idea is pivotal in discussions about the nature of consciousness and perception, particularly in understanding how different people might experience the same stimulus in varied ways. In the context of colour perception, qualia are especially significant because they highlight the internal, personal aspect of seeing colours, which cannot be fully captured by objective measures alone.

The way in which we understand colour is a window into how we perceive the world at large, which makes it all the more surprising that the category is so unstable diachronically (Söderqvist, 2017; Foley, 1997). Anna Wierzbicka (2006) explores this instability in her work on color semantics, highlighting the dynamic nature of color perception and its cultural underpinnings. She demonstrates that color terms undergo significant diachronic changes, with their meanings expanding, contracting, and merging at a higher rate than that of other lexical categories.

The constant evolution of culture influences how we describe colors and the words we colexify with them, reflecting broader shifts in our cultural landscape. This variability makes the study of color terms an intriguing area of inquiry, revealing insights into the dynamic evolution of language and its interplay with human cognition and cultural change. By understanding these processes, we can better appreciate the interconnectedness of our sensory experiences and linguistic expressions.

1.1 Colour Semantics

Colour term semantics has been at the centre of a heated debate around linguistic typological universals since the landmark study of Berlin and Kay (1969), which asserted that words for colours follow a universal evolutionary sequence across all languages. This study suggested that languages develop Basic Colour Terms, or BCTs (effectively monolexemic terms with reference to a primary colour; will be discussed in more depth later), in a strict order, starting with a binary distinction of terms for DARK/COOL and LIGHT/WARM, followed by RED, then one of YELLOW or GREEN, then both of them, followed by BLUE and subsequently BROWN, then, finally, all other colour terms (PURPLE, ORANGE, GREY, PINK). This particular universal can be justified somewhat more "sensibly" than other, more linguistically-oriented ones, due to the "panhuman neurophysiology of human vision" (Foley, 1997) — i.e. our vision systems are standard across all present-day peoples on Earth, excepting, of course, those with visual impairments.

As such, the cross-linguistic study of colour terminologies has emerged as a quintessential example to illustrate the influence of innate biological constraints on how we try to fit the world into neat linguistic categories. It's been assumed that if we can find such salient universals in the colour domain, then it's likely that we'll find comparable universals in various other domains (Levinson, 2022; Foley, 1997)

Naturally, Berlin and Kay's (1969) BCT theory has not existed without a platitude of valid criticisms and a selection of counterexamples levied towards it. However, before exploring these counterarguments, we will first recap how they define BCTs, as an important foundation for understanding the subsequent debates.

BCTs are defined as such:

1. BCTs are monolexemic; that is, their meanings are not predictable from the meaning of their parts (e.g. red is a BCT, but red-yellow is not)

2. BCTs are monomorphemic; meaning they are comprised of only one unit of meaning (e.g. blue, not *blueish*)

3. A BCT's signification is not included in that of any other colour term (e.g. mauve is a shade of purple)

4. Their application must not be restricted to a narrow class of objects (e.g. Irish *rua* generally describes the coppery reds of hair or rust, not reds in general)

5. BCTs must be psychologically salient. Indicators of this include (1) a tendency to appear first when various colour terms are elicited to describe something (2) consistency of reference among different informants, and (3) usage in the idiolects of all informants.

Despite the theoretical clarity provided by Berlin and Kay, numerous studies have identified variations and exceptions to their proposed universality of BCTs. For example, as highlighted by Wierzbicka (2006), the meanings of color terms can broaden, narrow, and merge over time, challenging the stability of BCTs across languages and cultures.

Even though Berlin and Kay's BCTs appear to be clearly defined, their list of BCTs, which includes PINK, and GREY seems to contradict their own rules, which may be the result of Anglocentric bias. Given that the aforementioned two colours can be made by just adding white to red and black, respectively, if we include these as BCTs, why would we not also include light blue, a colour that many languages do see as separate (e.g. *Azzurro* in Italian, голубой (*golubój*) in Russian).

One prominent critic of BCT theory, Stephen Levinson (2000), wrote extensively about the issues surrounding Berlin and Kay's (1969) assertions in his work *Yélî Dnye and the Theory of Basic Colour Terms*, discussing a language which seems to completely negate the universalist idea of BCT evolution.

Yélî Dnye (sometimes *Yele* or simply *Yélî*) is a language spoken by roughly 4,000 people on Rossel Island (*Yela*, to the locals), off the southeastern tip of Papua New Guinea. The language is tentatively classified as an isolate, with linguists failing to meaningfully connect it to any surrounding languages (see Levinson (2022) for more). Levinson's (2000) work on the language fundamentally undermines the universality of BCT theory, showing that existing colour terms in Yélî are not so "basic".

First of all, Yélî has no superordinate word for *colour*, relying on constructions like "U pââ ló nté?" [*Its body, what is it similar to?*], which may elicit any number of adjectival descriptions in addition to what we could consider "colour". Here *ló* is a general question marker, attached to *nté* 'like', generally used to mark comparisons or similes, leading to a consideration that asking for the colour of an object is just asking to compare it to something else. This reflects the idea that colour terms exist purely as a simple method to mark visual similarity to another object, rather than some innate property.

This absence of a superordinate colour term is accompanied by the relative absence of lexically distinct subordinate terms ("hyponyms") for individual colours. Analysing discourse, Levinson (2000) finds two major methods for deriving colour terms in Yélî, namely, Nominal Reduplication and Descriptive Phrases.

The former takes well-known nouns which have distinctive colours and reduplicates them to create an adjective that describes its colour. For example, *mtye* is a species of red parrot, and *mtyemtye* is the most frequently-used expression of "redness". There are similar constructions for other colours, like *mgîdî* "night" and *mgîdîmgîdî* "dark/black". Levinson (2000) notes that when answering a question like "U pââ ló nté?" the U pââ phrase is generally preserved. Thus, one can ask,

(1) *Yi 'nmeni u pââ ló nté ?*
DeicAnaph[1] bird$_{\text{SPEC}}$ his/her/its body what like ?
'That bird, what is its body/appearance like?'

And get the answer,

(2) *U pââ mgîdîmgîdî*
Its body black
'It is black in appearance.'

The second type of construction that creates color terms can be exemplified by *yi kuu yââ* "tree raw/uncooked leaf" and refers to fresh young (green) leaves. Unlike reduplicated terms, the entire descriptive color phrase usually follows the noun it modifies. However, some informants allow the descriptive phrase to precede the noun, as illustrated below. This indicates that the descriptive phrase, unlike the reduplicated forms, which tend to be adjectival, can be understood as a modifying compound noun.

(3) *u yi kuu yââ puku*
that tree raw/unripe leaf/roof book
'That green book'

Recall Berlin and Kay's (1969) criteria for defining a BCT. If we try to apply these criteria to the given Yélî expressions, we can start to question whether there are any BCTs in the language at all. All of this is to say that BCT theory is not as universal as its proponents would argue. In fact, its universality can be called further into question in the study of colexification.

[1] Deictic Anaphoric. Indicates that "Yi 'nmeni" functions both to point out a specific bird and to refer back to the previously mentioned or contextually understood bird.

1.2 Colexification

The term “colexification”, coined by Alexandre François (2008), describes instances where a single word form expresses multiple semantic senses, regardless of whether this arises from polysemy or homophony. Common examples of colexification cross-linguistically come from body part terms — Spanish *dedo* can refer to both a finger and a toe, while Irish *cos* colexifies the concepts for *foot* and *leg*. English *dislexifies* both of these examples, employing separate lexemes for all four concepts.

François (2008) delineates between *strict* and *loose* colexifications, with the former requiring the existence of a single lexeme used synchronically that denotes two distinct semantic senses, while the latter may include diachronic links and those based on composition or derivation. To illustrate this point, the examples above of *dedo* and *cos* are strictly colexified as they each represent two conceptual meanings in synchrony, while instances where the connections between meanings are less direct, such as with historical semantic shifts or where the meanings are linked through derivational processes, would fall under the umbrella term of loose colexification. This distinction helps to clarify the varying degrees of relationship between concepts that share a lexical item and offers a more nuanced understanding of how meanings can be interconnected across languages and time.

One of the major historical reasons for the existence of so many colexified concepts is a linguistic cycle, originally posited by Matthias Urban (2015), which describes the situation in which a word originally has meaning A, then gradually adopts meaning B as well, generally through broadening or metaphor, before often losing the original meaning. For example from Söderqvist (2017), Spanish *alcalde* was borrowed from Arabic *qāḍī*, ‘judge (in Islamic law)’ (stage A), then later broadened and colexified to mean ‘an official who is magistrate and mayor’ (stage A∼B) before eventually losing its original meaning to come to only mean ‘mayor’ today (stage B). Many words exist in the A∼B state for an extended time, leading us to consider them synchronically colexified. This type of language change is somewhat reminiscent of the Jespersen cycle (Jespersen, 1917), as well as various other cyclic behaviours ever-present throughout the field of linguistics (see Van Gelderen (2011) for more).

This example of the linguistic journey from a specific judicial role to a more generalised municipal position demonstrates how social and political changes can also drive linguistic evolution. Overall, colexification is a vital concept in linguistics that bridges the gap between semantics, typology, and historical linguistics. It provides a framework for understanding how different languages encode meaning and how these meanings can evolve over time. As research in this area continues to grow, it promises to shed further light on the complex interplay between language, thought, and culture.

1.3 Research Question

With the theoretical background out of the way, we turn to the central research question of this study.

> What patterns can be revealed in the domain of colour term colexification through the exploration of linguistic databases and how does this affect our understanding of universalist approach to linguistic colour theory?

Before tackling this question, let’s firstly look at some work previously carried out in this field.

2 Related Work

Kajsa Söderqvist (2017) takes a historical, etymological approach to defining colexifications between colour terms, analysing ten Sino-Tibetan and ten Indo-European languages. This methodology leads to the discovery of a far broader set of potential diachronic colexifications, many of which would never be found in the synchronic approach of CLICS (Rzymski et al., 2020) and others.

Söderqvist (2017) includes 10 of the 11 BCTs defined by Berlin and Kay (1969), excluding PINK as it didn't commonly exist as a colour term in Europe until the 17th or 18th century, making it irrelevant for an etymological study. The dataset used was created by the author herself and has been scraped from various etymological dictionaries — mainly STEDTS (Sino-Tibetan Etymological Database and Thesaurus) for the Sino-Tibetan languages and Brill Dictionary Online for the Indo-European ones. Söderqvist (2017) also defines 8 semantic categories to classify the lexical meanings of discovered colexifications. These categories are: The Physical World, The Body, Sense Perception, Food and Drink, Animals, Basic Actions & Technology, Agriculture & Vegetation, and Colour. We will attempt to alter these categories as little as possible so that they can be repurposed in the following analysis in order to maintain cohesion between different studies in this field. Later we will explore whether they fit our data as well as they do in this case, and what shortcomings might befall them.

Annika Tjuka's (2024) recent study on body-object colexifications forms a methodological basis for this work, providing an excellent example of a workflow for extracting, analysing, and visualising colexification data from CLICS^3. The author's self-stated goal is to examine the frequency, distribution, and cognitive relations of body-object colexifications, which she does exceptionally through her implementation of a rigorous methodology. This methodology of data extraction, analysis, processing, and visualisation can be adapted for our purposes. The fact that the author's code is available online is of great use for understanding how to work with data in the CLDF format as intended. Tjuka (2024) finds many widespread body-object colexifications, and many more that only appear in a few languages. The author makes it clear as well that the large-scale approach has various limitations — in particular, it's discussed how areal patterns of loose colexifications can get obfuscated in the coarse grid of an automated methodology. This limitation emphasises the persistent need for more detailed investigations that go beyond the analysis of previously-available aggregated datasets.

Investigating areal patterns in more depth than Tjuka (2024), Segerer and Vanhove (2021) explore RefLex (Segerer & Flavier, 2011-2023), a huge cross-linguistic lexical database of African languages in their study on areal patterns of colour colexifications in the languages of Africa. The authors discover several previously undetected areal colexification patterns, proving that it is possible to extract useful and novel information from large databases not specifically built for typological purposes. Multiple types of lexico-semantic phenomena were identified through this research. Most notable for our study, it was shown that there would be shared colexification within a Sprachbund (e.g. RIPE and RED) as well as randomly distributed colexification (e.g. UNRIPE and GREEN) for two seemingly related (or opposite) concepts.

3 Methodology

3.1 Definition

For the purposes of this study, we minorly adapt François' (2008) definition of colexification to include only *strict* colexifications.

> A language is said to *colexify* two functionally distinct senses if, and only if, a single lexical item in a given language simultaneously represents both meanings synchronically. The term is intended to be used descriptively, rather than prescriptively, and is neutral with respect to semantic and historical interpretation.

This definition emphasises the following criteria:

1. The meanings involved must be considered separate and distinct, rather than merely different shades or aspects of the same concept

2. The distinct meanings must be distinguishable based on contextual cues, allowing for correct interpretation in various situations

3. The phenomenon must occur synchronically, meaning that the single lexical item carries these multiple meanings at the same time, rather than being a result of historical or diachronic semantic development

Unlike Söderqvist's (2017) etymological, diachronic approach, we only consider colexifications used in synchrony. Though a diachronic study would be interesting, we are restricted by the methodology of how the data was gathered for CLICS, who define colexifications as above.

3.2 Database: CLICS^3

Colexifications can be algorithmically scraped from extensive collections of lexical data, particularly from (often historical) multilingual wordlists, which translate a set of concepts into various target languages. This approach has contributed to the expansion of the Database of Cross-Linguistic Colexifications (CLICS) (Rzymski et al., 2020) in its recent versions; an invaluable resource for this project.

The Third Database of Cross-Linguistic Colexifications (CLICS^3) (Rzymski et al., 2020) is a database and framework built to interactively represent cross-linguistic colexification patterns. This third release of CLICS addresses various shortcomings of the first version, and more than doubles the size and scope of version two (see Rzymski et al. (2020) for more).

Databases like CLICS are essential for quantitative approaches in linguistics because they provide structured, standardised, and accessible data that can be analysed using statistical and computational methods. By aggregating vast amounts of cross-linguistic data, CLICS enables researchers to uncover patterns and correlations that would be difficult to detect through qualitative methods alone. These quantitative analyses can lead to insights into language universals, typological tendencies, and the dynamics of language change and contact.

All of CLICS' data is stored in the Cross-Linguistic Data Format (CLDF), defined by Forkel et al. (2018) in order to standardise cross-linguistic data, which decouples the development of methods from individual datasets, allowing them to be more broadly applied to any given data. By adhering to CLDF standard, CLICS can easily integrate diverse datasets from various linguistic sources into a unified framework, facilitating the creation of a readily-available and useful dataset which can be explored simply through algorithmic methods.

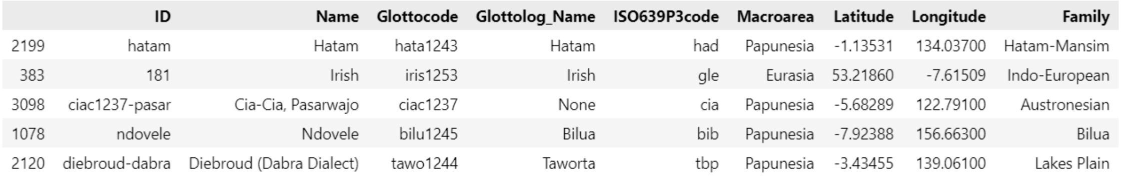
The main CLICS dataset is freely available online, represented as a *graph* and saved in the Geography Markup Language (GML) format. In the field of computer science, a graph is an abstract data structure, consisting of (weighted) nodes and edges and showing pairwise relationships between objects. In this case, semantic concepts are represented as nodes and their colexifications as edges, with the nodes weighted by occurrence frequency and the edges by cross-language frequency.

3.3 Language Families

In order to manageably analyse the CLICS data in a meaningful way, we have chosen to study and compare preferences for colour term colexification in 6 language families, namely, Indo-European, Austronesian, Sino-Tibetan, Nakh-Daghestanian, Atlantic-Congo, and Nuclear Trans-New Guinea.

These broad language families are some of the largest and most well-attested in cross-linguistic databases, meaning our conclusions can be stronger and based on more data. An explicit treatment/discussion of all the languages surveyed is outside of the scope of the study, and we instead make generalisations to broader language families to give a macro-overview of some key differences. There is, of course, a trade-off of granularity underpinning this decision, as more insightful results could possibly be gleaned from looking at individual languages. However, due to the sparsity with which languages tend to employ colexifications, a broad, top-level survey will give a higher chance of seeing statistically significant results per family.

While the languages chosen to be included in language families certainly can be debated, we simply use the CLICS definitions, which come from Glottolog (2021), shown in the LanguageTable of the main SQLite database (sample shown below).



	ID	Name	Glottocode	Glottolog_Name	ISO639P3code	Macroarea	Latitude	Longitude	Family
2199	hatam	Hatam	hata1243	Hatam	had	Papunesia	-1.13531	134.03700	Hatam-Mansim
383	181	Irish	iris1253	Irish	gle	Eurasia	53.21860	-7.61509	Indo-European
3098	ciac1237-pasar	Cia-Cia, Pasarwajo	ciac1237	None	cia	Papunesia	-5.68289	122.79100	Austronesian
1078	ndovele	Ndovele	bilu1245	Bilua	bib	Papunesia	-7.92388	156.66300	Bilua
2120	diebroud-dabra	Diebroud (Dabra Dialect)	tawo1244	Taworta	tbp	Papunesia	-3.43455	139.06100	Lakes Plain

Figure 3.1: Sample of Language Metadata from CLICS

3.4 Data Processing

The data processing pipeline designed for this study firstly reads and parses the network of colexification data from CLICS stored in a GML file, as well as additional language family data from an SQLite database. The entire CLICS dataset is represented as a graph with 2919 nodes (concepts) and 4228 edges between them (colexifications). Though we would like to analyse all 11 of Berlin and Kay's (1969) BCTs, CLICS only records colexifications for seven of them, namely RED, YELLOW, BLUE, WHITE, BLACK, GREEN, and GREY, with ORANGE, PURPLE, and BROWN having no attested colexifications and PINK not even being listed as a concept. Reducing the original dataset to just the concepts referring to colour terms or those directly linked to them, we get a graph with 35 nodes and 60 edges. This can be visualised as in Figure 3.2. For readability, only edges that link to at least one colour word are shown (i.e. FIRE and FLAME also colexify with each other as well as with RED but this isn't particularly important knowledge for our study). After creating subgraph visualisations for each colour and language family, the system computes and prints summary statistics for them too.

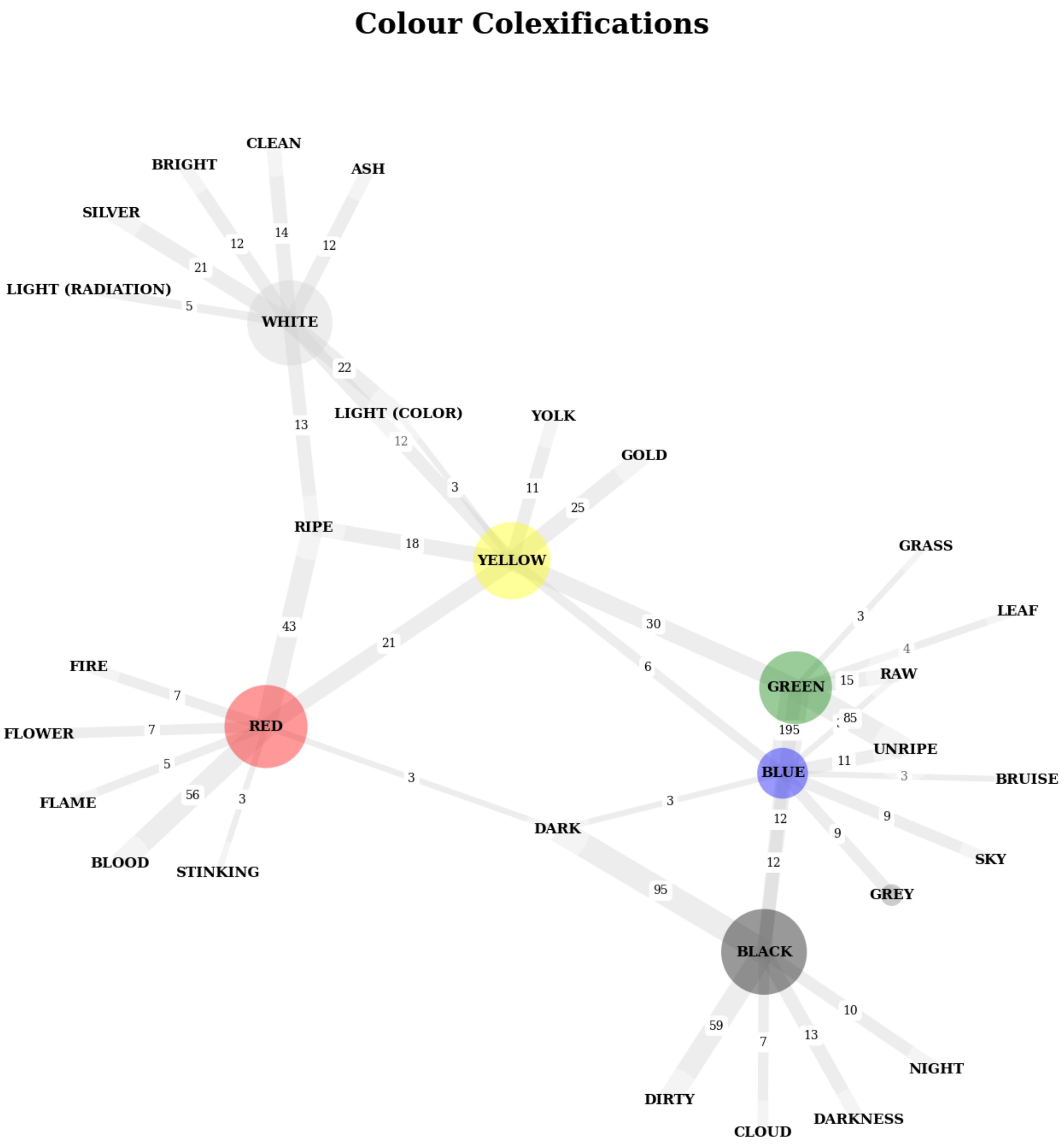


Figure 3.2: Full graph of all colour colexifications in CLICS. Colour nodes are scaled by the frequency with which a lexeme bound to the concept exists in a language

4 Results

4.1 General Observations

The most common colour-colour colexification represented in the data is BLUE/GREEN, attested in 195 languages. As also demonstrated by Söderqvist (2017), this is to be fairly expected, given the frequent connection between the terms and their overarching macro-category GRUE.

GREY only colexifies with BLUE, not with BLACK or WHITE, which may well demonstrate that it does only appear later in the evolution of BCTs, as suggested by Berlin and Kay (1969). YELLOW and BLUE are the colour terms most colexified with other colour terms, both colexifying with 4 other BCTs.

The study revealed no terms for animals colexified with any colours. This shows a stark difference from Söderqvist's (2017) diachronic study, which revealed various etymological relations between animals and colours (e.g. WORM, MAGGOT, and MONKEY with RED, PIGEON with BLUE, and BEAVER and ELK with BROWN).

All of the attested colexifications are shown in Table 4.1. In general, the semantic categories are less well-populated in these results than in Söderqvist's (2017) diachronic study. The comparative lack of colexifications makes the semantic categorisation less meaningful, but some language families do seem to have a slight preference for one of them. For example, the major colexifications in Nakh-Daghestanian languages are GREEN/UNRIPE and YELLOW/YOLK, which could both be placed in the Food & Agriculture category.

Table 4.1: Colexified Concepts Organised by Semantic Category

Physical World	**The Body**	**Sense Perception**	**Food & Agriculture**	**Actions & Technology**
ASH	BLOOD	BRIGHT	FLOWER	FIRE
CLOUD	BRUISE	CLEAN	GRASS	FLAME
DARKNESS		DARK	LEAF	
GOLD		DIRTY	RAW	
LIGHT (RADIATION)		LIGHT (COLOUR)	RIPE	
NIGHT		STINKING	UNRIPE	
SILVER			YOLK	
SKY				

4.2 Colexifications by Language Family

Table 4.2: Total Attested Colour Colexifications by Language Family

Language Family	**RED**	**YELLOW**	**BLUE**	**WHITE**	**BLACK**	**GREEN**	**GREY**
Indo - European	6	18	20	11	5	40	4
Austronesian	10	7	33	5	32	37	0
Sino - Tibetan	5	14	31	17	29	32	2
Atlantic - Congo	0	0	1	0	1	1	0
Nakh - Daghestanian	1	9	7	0	1	29	0
Nuclear Trans New Guinea	47	3	3	8	10	5	0

Above is a summary table of all of the attested colexfications per colour term, split by language family. We can immediately see some families that stand out, for example Atlantic-Congo, which appears to have almost no attested colour colexifications in this data. This will be explored further in a per-family basis later in this section. Below, we see another table which demonstrates the ratio of colour-based colexifications in a language family in comparison to the total number of colexifications. Evidently, Nakh-Daghestanian languages seem to colexify terms at a higher rate than the other families, but also seem to have relatively fewer colexifications involving colour terms.

Table 4.3: Ratio of Colour to Total Colexifications by Language Family

Language Family	**Colour Colexifications**	**Total Colexifications**	**Ratio % (Colour/Total)**
Indo - European	84	9938	0.85
Austronesian	87	4835	1.8
Sino - Tibetan	95	3498	2.72
Atlantic - Congo	2	3235	0.06
Nakh - Daghestanian	41	12092	0.34
Nuclear Trans New Guinea	68	1950	3.49

4.2.1 Indo-European

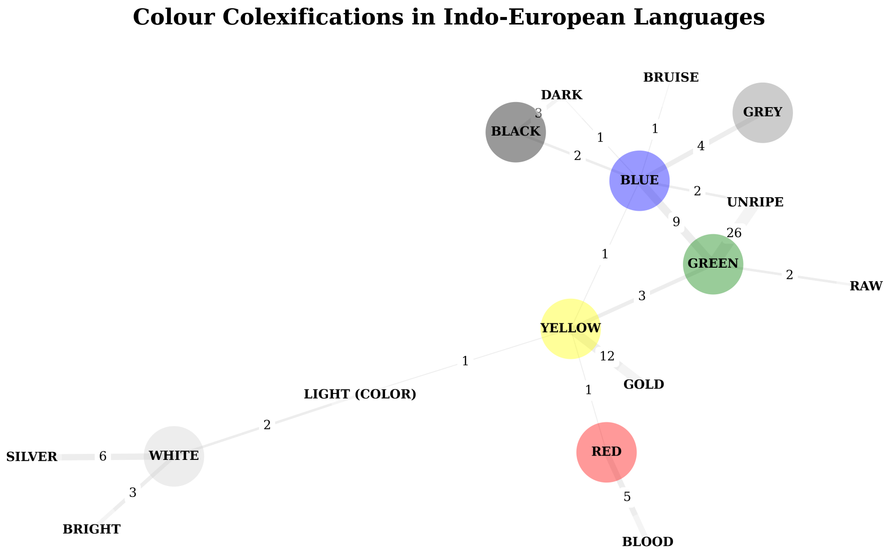


Figure 4.1: Graph of Colour Colexifications in Indo-European Languages

The most frequent colour colexification in Indo-European languages is GREEN/UNRIPE, attested in 26 of the 266 IE languages in the CLICS dataset. GOLD and SILVER colexified 12 and 6 times with YELLOW and WHITE, respectively, showing an interesting frequency bias between two apparently similar colexifications. When analysing colour-colour colexifications, BLUE/GREEN is the most frequent, again reflecting a preference for GRUE over other colour pairings. Both YELLOW and WHITE colexify with LIGHT (in colour), while both BLUE and BLACK colexify with dark, indicating a perception that blue shades are generally darker and yellow shades lighter. This pattern of colexification provides insights into the cultural and environmental factors that shape language. For instance, the prevalence of GREEN/UNRIPE may suggest an agrarian focus where the ripeness of crops is significant. This particular colexification is only common in the Indo-European and Nakh-Daghestanian families.

4.2.2 Austronesian

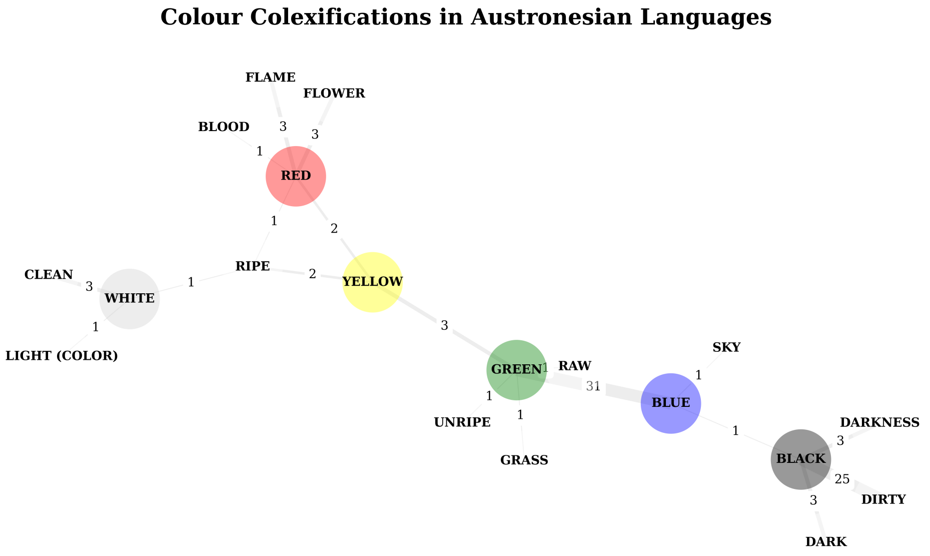


Figure 4.2: Graph of Colour Colexifications in Austronesian Languages

Austronesian languages are spoken widely across Maritime Southeast Asia, from Madagascar to Hawai'i, and include Malagasy, Indonesian, Malay, Tagalog, and Samoan. These languages record a total of 4835 colexifications, with 87 attested colour-based samples. BLUE/GREEN is again highly attested by 31 languages. Another frequent colexification is BLACK/DIRTY, which is completely unattested in all other language families, except for one instance in a Nuclear Trans New Guinea language, which may be areally related to the Austronesian samples. Interestingly, there is an imbalance here too, with WHITE/CLEAN only appearing in three languages. Aside from the two common colexifications, all others appear infrequently, with only one to three instances for colexifications like RED/FLAME and YELLOW/GREEN. GREEN/UNRIPE is practically unattested, appearing in only one language.

4.2.3 Sino-Tibetan

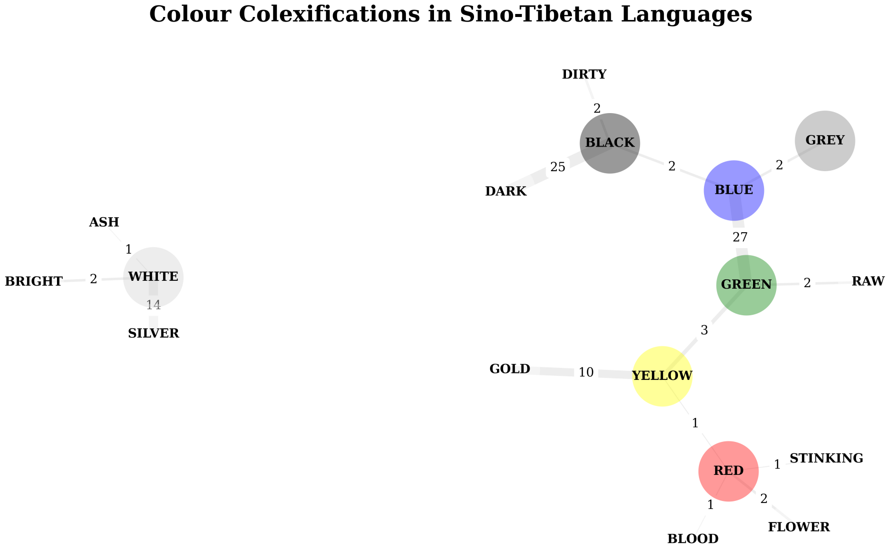


Figure 4.3: Graph of Colour Colexifications in Sino-Tibetan Languages

Sino-Tibetan languages are spoken across mainland China, the Himalayas, and Myanmar. They have the second highest ratio of colour-based colexifications of the language families in this study. The Sino-Tibetan languages also show a high amount of BLUE/GREEN colexifications, with 27 language instances. Like Indo-European, GOLD and SILVER are frequently colexified with YELLOW and WHITE, respectively. However, these colexifications show the opposite frequency bias, with 14 languages colexifying WHITE/SILVER and 10 colexifying YELLOW/GOLD. These languages also have a strong tendency towards the BLACK/DARK colexification, which is an interesting contrast to Austronesian languages tendency towards BLACK/DIRTY. They are also the only family other than Indo-European to colexify BLUE/GREY.

4.2.4 Atlantic-Congo

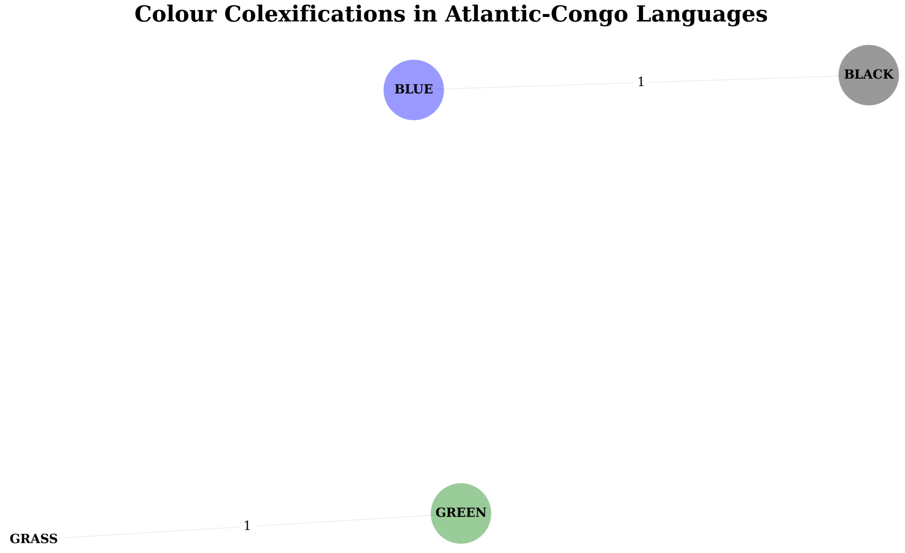


Figure 4.4: Graph of Colour Colexifications in Atlantic-Congo Languages

The Atlantic-Congo languages form the core of the proposed broader Niger-Congo family and stretch all the way across the width of the African continent. Only two colour-based colexifications were found of the 3235 total colexifications in the 155 Atlantic-Congo languages present in CLICS. These results are surprising and exist in sharp contrast to those of Segerer and Vanhove (2021), who found a platitude of examples of colexifications in Atlantic-Congo languages, including a tendency towards RED/RIPE and GREEN/UNRIPE colexifications. This may indicate a chasm of missing colexifications in the CLICS data, a point mentioned by the authors themselves as well (Segerer & Vanhove, 2021). BLUE and BLACK are colexified, which is interesting, as this is not so common in the rest of the language families, only appearing one or twice in some of them. However, given the questionable state of the data, no more can meaningfully be interpreted from these results.

4.2.5 Nakh-Daghestanian

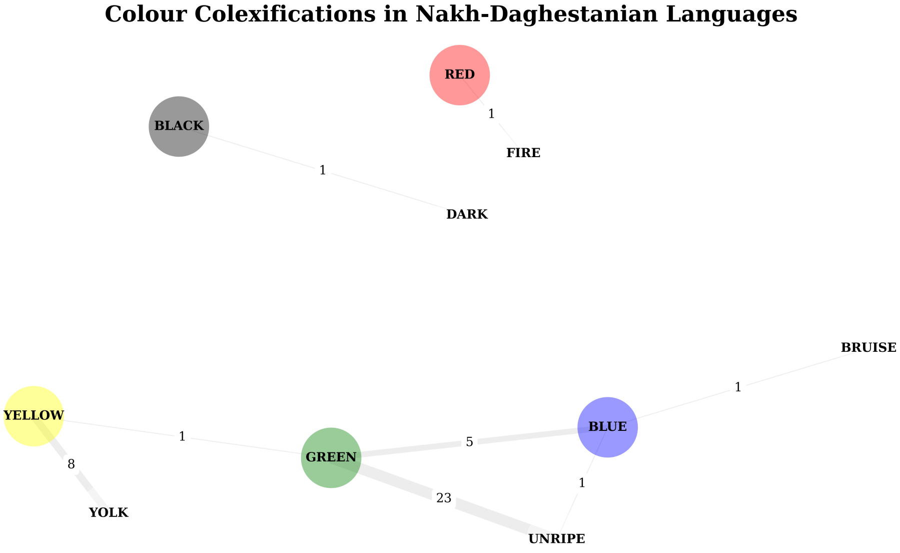


Figure 4.5: Graph of Colour Colexifications in Nakh-Daghestanian Languages

The Nakh-Daghestanian language family, spoken mostly in the northwest Caucasus by the Caspian Sea, has the most attested colexifications of any family in CLICS, while only encompassing 107 languages. Strangely, though, it has a very low rate of colexification for colour terms, with only 41 of its 12,092 being related to a BCT. The most frequent colour colexification here is GREEN/UNRIPE, with 23 instances: a similar amount to the Indo-European languages. There are also 5 instances of BLUE/GREEN colexification. The only other notable colexification is YELLOW/YOLK, which is unattested in all other language families. To tentatively proffer an additional hypothesis for the prevalence of the GREEN/UNRIPE colexification, the consumption of food is one of the few aspects of our lives where colour is critical for survival. Ripeness is an obvious indicator, but important too is species identification, and indicators of e.g. poisonous or rotten food. There are also a plethora of colour descriptors beyond BCTs which originate from food terms (chocolate, cream, mint, etc.).

4.2.6 Nuclear Trans New Guinea

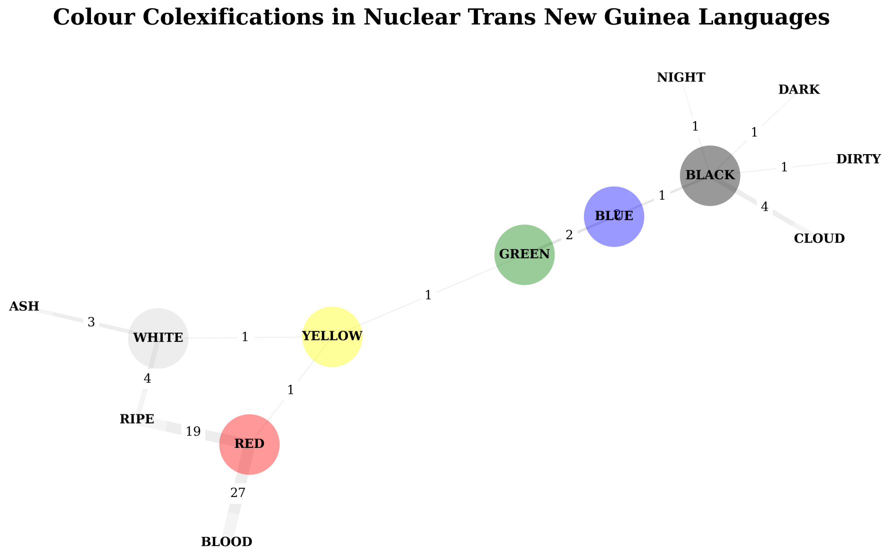


Figure 4.6: Graph of Colour Colexifications in Nuclear Trans New Guinea Languages

The Nuclear Trans New Guinea languages, of which there exist 462 in this dataset, seem to show a huge preference for colexifications involving RED, with it being joined to BLOOD and RIPE far more frequently than in any other family. Thanks to these, the NTNG family has the highest ratio of colour-based colexifications of the six families in this study, with 68 of its 1950 colexifications being related to colour terms. Other than the RED colexifications, there is little of note here. CLOUD is colexified with BLACK 4 times and never with WHITE, raising a notion that only rain-filled clouds are noteworthy enough to be lexically associated with a colour.

5 Conclusions

This study presents a data-driven analysis of the various ways in which different language families colexify their basic colour terms. It reveals significant cross-linguistic variations and patterns that provide insights into linguistic and cultural perceptions of colour.

In the Indo-European languages, the most common colour colexification is GREEN/UNRIPE, suggesting a cultural emphasis on agriculture and the ripeness of crops. This colexification pattern also appears prominently in Nakh-Daghestanian languages, indicating a possible shared cultural or environmental influence.

Austronesian languages display a notable preference for BLUE/GREEN colexifications, consistent with the maritime environments these languages are often associated with. The BLACK/DIRTY colexification, prevalent in Austronesian languages but rare elsewhere, highlights unique cultural perceptions of cleanliness and colour.

Sino-Tibetan languages show a strong tendency towards BLUE/GREEN and BLACK/DARK colexifications, reflecting perhaps a different set of environmental or cultural associations compared to other language families. The frequent colexification of WHITE/SILVER and YELLOW/GOLD in these languages points to distinct symbolic uses of these colours.

The Atlantic-Congo language family exhibits surprisingly few colour-based colexifications, in sharp contrast to previous studies suggesting richer lexical associations. This discrepancy underscores the need for more comprehensive data to fully understand colour perceptions in these languages.

Overall, this study contributes to the field of lexical typology, providing a literature review, a methodological workflow, and a comprehensive analysis of available synchronic data on colexifications.

The broad scope of this study leaves the potential for a lot of missing nuance. Future work in this field should aim at carrying out a more extensive survey of more languages and language families. The nature of the available data is also far from perfect, and more work should be done to aggregate available data into more comprehensive datasets of cross-linguistic colexifications.

Resources

Abbreviations

CLICS	Database of Cross-Linguistic Colexifications (Rzymski et al., 2020)
BCT	Basic Colour Term (Berlin & Kay, 1969)
CLDF	Cross-Linguistic Data Formats (Forkel et al., 2018)

Extras

The code written to extract, analyse, and visualise colour colexifications is available in a GitHub repository at https://github.com/joshbrook/colour-colex

Subgraphs divided by colour term are available in the Appendix.

References

Berlin, B., & Kay, P. (1969). *Basic color terms: their universality and evolution.* Berkeley: California University Press.

Foley, W. (1997). *Universalism: Innate Constraints on Mind: Color.* Blackwell Publishers.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., ... Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, *5*(1). doi: 10.1038/sdata.2018.205

François, A. (2008). *Semantic maps and the typology of colexification: Intertwining polysemous networks across languages.* doi: 10.1075/slcs.106.09fra

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). Glottolog 5.0. *Leipzig: Max Planck Institute for Evolutionary Anthropology.*. doi: 10.5281/zenodo.596479

Jespersen, O. (1917). *Negation in English and other languages.* Retrieved from `http://ci.nii.ac.jp/ncid/BA31919234`

Kanai, R., & Tsuchiya, N. (2012). Qualia. *CB/Current biology*, *22*(10), R392–R396. doi: 10.1016/j.cub.2012.03.033

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by.* University of Chicago Press.

Levinson, S. (2000). Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, *10*, 3 - 55. doi: 10.1525/jlin.2000.10.1.3

Levinson, S. (2022). *A grammar of yélî dnye: The papuan language of rossel island.* De Gruyter Mouton. doi: doi:10.1515/9783110733853

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., … List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, *7*(1). doi: 10.1038/s41597-019-0341-x

Segerer, G., & Flavier, S. (2011-2023). *RefLex: Reference Lexicon of the Languages of Africa.* Retrieved from `https://reflex.cnrs.fr/` (Paris, Lyon)

Segerer, G., & Vanhove, M. (2021). Areal patterns and colexifications of colour terms in the languages of Africa. *Linguistic typology*, *26*(2), 247–281. doi: 10.1515/lingty-2021-2085

Söderqvist, K. (2017). *Colexification and semantic change in colour terms in Sino-Tibetan and Indo-European languages* (BSc Thesis, University of Lund, Centre for Languages and Literature). Retrieved from `https://www.lunduniversity.lu.se/lup/publication/8905470`

Tjuka, A. (2024). Objects as human bodies: cross-linguistic colexifications between words for body parts and objects. *Linguistic typology*, *0*(0). doi: 10.1515/lingty-2023-0032

Urban, M. (2015). *Lexical semantic change and semantic reconstruction.* doi: 10.4324/9781315794013.ch16

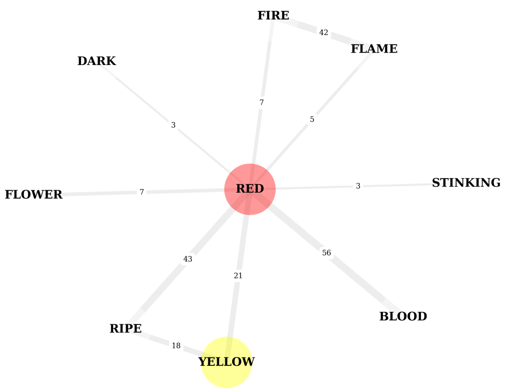
Van Gelderen, E. (2011). *The Linguistic Cycle: Language Change and the Language Faculty.* Oxford University Press. doi: 10.1093/acprof:oso/9780199756056.001.0001

Wierzbicka, A. (2006). The semantics of colour: A new paradigm. *Progress in Colour Studies*, *1*, 1-24.

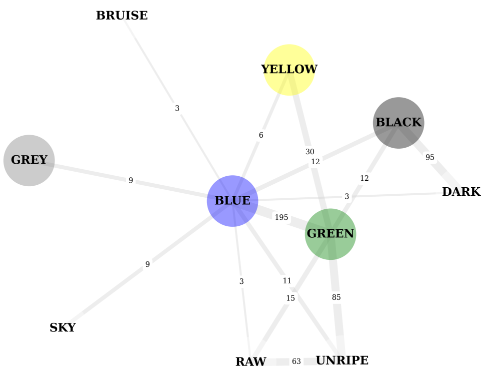
Wierzbicka, A. (2008). Why there are no 'colour universals' in language and thought. *Journal of the Royal Anthropological Institute*, *14*(2), 407–425. doi: 10.1111/j.1467-9655.2008.00509.x

Appendix

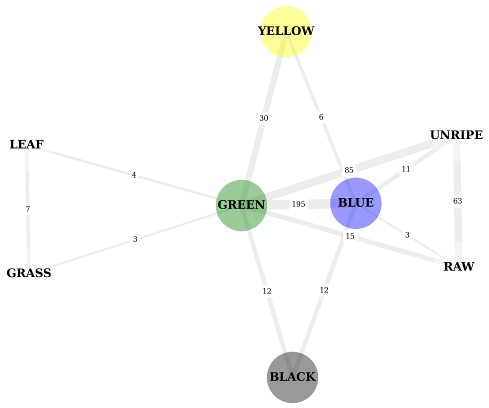
Subgraph RED



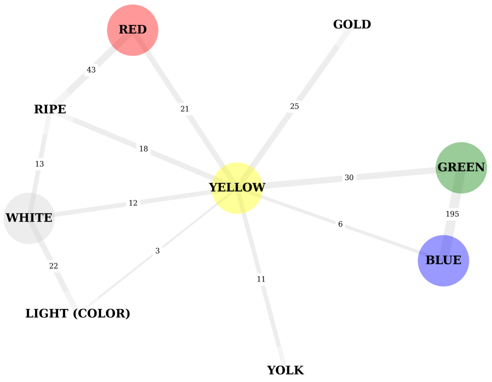
Subgraph BLUE



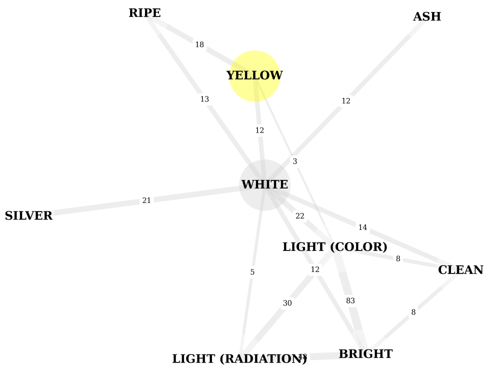
Subgraph GREEN



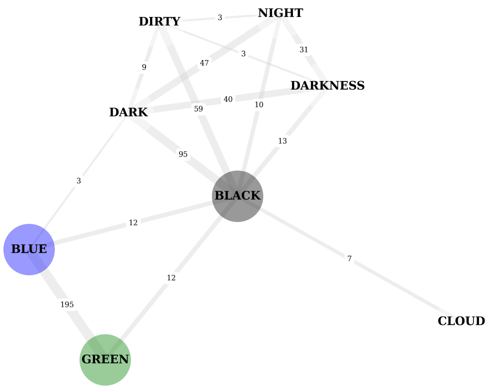
Subgraph YELLOW



Subgraph WHITE



Subgraph BLACK


DIRTY
3
NIGHT
3
31
9
47
DARKNESS
DARK
59
40
10
13
95
3
BLACK
12
BLUE
7
12
195
CLOUD
GREEN