Received: 6 May 2023 | Revised: 5 July 2023 | Accepted: 22 August 2023

DOI: 10.1002/col.22899

RESEARCH ARTICLE

COLOR RESEARCH AND APPLICATION WILEY

Quantifying basic colors' salience from cross-linguistic corpora

Antoni Brosa-Rodríguez | **M. Dolores Jiménez-López**

Universitat Rovira i Virgili, GRLMC-Research Group on Mathematical Linguistics, Tarragona, Spain

Correspondence
Antoni Brosa-Rodríguez, Universitat Rovira i Virgili, GRLMC-Research Group on Mathematical Linguistics, Av. Catalunya 35, Tarragona 43002, Spain.
Email: antoni.brosa@urv.cat

Funding information
Marti i Franques Research Fellowship Programme, Grant/Award Number: 2019PMF-PIPF-99-; GRLMC - Research Group on Mathematical Linguistics, Grant/Award Number: 2021-SGR-00032 -

Abstract

A corpus-based quantitative assessment of Berlin and Kay's proposal is presented. We refine the Basic Color Terms hierarchy proposed by Berlin and Kay, through the concept of salience. A cross-linguistic study with 57 different languages and 136 different linguistic corpora has been conducted. This study uses KonText tool and the corpora included in it. The color labels in different languages have been obtained using a unified methodology from PanLex. We have obtained an individual hierarchy for each of the languages analyzed, as well as a general hierarchy that captures the universal trend. Results show that there is a close relationship between the evolutionary stages in the Berlin and Kay proposal and their frequency in our corpora study, which we could also relate to Zipf's Law. The only color that we certify behaves differently compared to such a proposal is *yellow*. The main advantage of our approach compared to previous corpora studies is taking into account the anglocentric bias by using a representative typological set of different languages from the world.

KEYWORDS

basic color terms, color metrics, linguistics, terminology

JEL CLASSIFICATION

Z13, Y10

1 | INTRODUCTION

This paper provides new data to complement the study of the Basic Color Term (BCT) hierarchy proposed by Berlin and Kay[1] in the field of Universals of Language. Our main interest is to obtain frequencies in order to calculate the salience of the different basic color terms. The concept of salience is related to the frequent activation (either in production or in comprehension) of an element in a whole linguistic community; the more frequent the occurrence, the more salient the term.[2] Our hypothesis is that the salience of colors is directly proportional to their chronological appearance in the different languages. That is, for any color, the sooner appearance, the higher use.

We confirm Berlin and Kay's proposal and refine their hierarchy with our salience data. In the cases in which Berlin and Kay propose different colors in the same level, we assess whether any of the colors are used more frequently. Moreover, we analyze the preference for different synonyms in a given language for a single English color term (if any).

In the last thirty years, there have been a lot of studies that review BCT proposal from different perspectives. Most

of those studies focus their attention in single languages.[3–5] The use of questionnaires is the most frequent technique used in this literature.[6,7] Important advances in this field have been achieved. However, as far as we know, there is a lack of quantitative cross-linguistic studies. With our proposal, we want to offer one more way of getting to know the behavior of colors words contributing in this way to the progress in the knowledge in this topic.

In this paper, we approach the phenomenon through linguistic corpora with the aim of offering a unified and holistic analysis of colors in real texts in 57 different languages, shown in Table A1 in the Appendix A. This analysis will allow us, on the one hand, to obtain these data on the use of languages for which we do not yet have hierarchies and, on the other hand, to offer these data as a complement to other experimental approaches. A benefit of our corpus approach (studying multiple languages, many of them with few resources) is the efficiency and cost-effectiveness of the study.

Besides the theoretical interest of the BCT hierarchy we highlight its practical and industrial applications.[8]

Berlin and Kay's proposal is an implicational hierarchy within the studies of universals in language and linguistic typology. Their theory "maintains that the world's languages share all or part of a common stock of color concepts and that terms for these concepts evolve in a constrained order".[9]

In order to understand Berlin and Kay's proposal it is important to say few words about language universals. In the 1960s, there was a revolution in linguistics studies thanks to the foundational publication of Greenberg et al.[10] on universals in language. The main aim was to find common features in all languages of the world.[11]

These universals were (or are) formulated by comparing a linguistic element in different languages of the world.[12–14] Since it is not possible to review all languages due to lack of data, a balanced selection is reviewed, as would be done in any similar study.[15–17]

Despite the fact that the universals presented usually contain exceptions,[18,19] their usefulness remains valid because of their broad predictive power.[20,21] For example, it is of great value for linguistics to know that if a language has the direct object after the verb (such as English or Spanish), it will use prepositions.

On the other hand, if a language has the direct object before the verb (like Basque or Japanese), it will use postpositions.[22–25] This is what is known as implicative universal,[15,26,27] which uses the "if-then" structure. In Croft's words[28]: "Implicational universals differ from unrestricted universals in that they do not claim that all languages belong to one type. Instead, they describe a restriction on logically possible language types that limits linguistic variation but does not eliminate it".

A major finding was what is known as an implicational hierarchy,[29,30] which is basically a sum of implicational universals. What is really interesting about this structure is that, as in any hierarchy, the order of the constituents is not random. That is to say, if a language contains an element of this scale, it will also have all the elements to its left, without exception. In addition, we should point out another very relevant feature in relation to implicational hierarchies: their predictive power (based on the prediction of the implications described above) is multiplied exponentially. This evolution can be either implementation or downscaling, but it will always occur in an orderly and predicted, non-random manner.

From a diachronic perspective, that is, the evolution of languages, it is possible to see how elements of the hierarchy have increased or decreased. This can be seen in the shift from Latin to Romance languages, for example. Implicational hierarchies can also be used to understand intralinguistic variation in a language, since different dialects of a language, for example, can be found at a different level of the hierarchy, but not at any point in the hierarchy.[31]

By way of example, we can highlight three of the four most famous hierarchies in typology, according to Corbett[31]: the accessibility hierarchy,[32] the animacy hierarchy[33] and the agreement hierarchy.[34] The fourth that remains to be mentioned is semantic and is the one we are concerned with in this paper, about colors and proposed by Berlin and Kay[1] as shown in Figure 1.

All the basic and fundamental aspects of this proposal and its historical evolution that have already been well summarized and explained by Jameson and Webster[8] (in a purely literature review paper) will not be dealt with in depth in this paper.

The first aspect to discuss is the distinction between colors that appear in the hierarchy and those that could appear, but do not (as they are not considered basic). To establish the distinction between colors that are considered basic and those that are not, Berlin and Kay use different linguistic criteria. Among them, we should highlight:

1. Not to depend on a hyperonym. That is to say, not to include in the meaning of that color being a "type of" (as happens with *navy blue*).
2. Not to be modified by morphemes (as is the case with *brownish*).

$$\left\{\begin{matrix}\text{white}\\\text{black}\end{matrix}\right\} < \text{red} < \left\{\begin{matrix}\text{green}\\\text{yellow}\end{matrix}\right\} < \text{blue} < \text{brown} < \left\{\begin{matrix}\text{purple}\\\text{pink}\\\text{orange}\\\text{grey}\end{matrix}\right\}$$

FIGURE 1 Implicational hierarchy of basic colors, according to Berlin and Kay.

3. That the extension of the color is general to any object and not restricted to a specific context (as with *blond*).
4. It must be monolexemic, that is, not composed of more than one lexeme (as is the case with *navy blue*).
5. That they are salient colors. That is, that their frequency of use or recognition is very high and not anecdotal (in relation to corpora), in any population group. Or that their identification appears in the first positions or within a few seconds (in relation to psycholinguistic tests). That is, presumably, it will be much more common and frequent to find *black* rather than *turquoise* in a text. Or, in the case of tests, the label *black* will be identified earlier and in less time than *turquoise* in a Munsell palette (if the latter color is identified at all).

This last criterion, *salience*, is the one that interests us in our study. In color research, it has traditionally been used to determine which colors are basic and which are not. Normally, in order to determine the evolutionary stages, the repertoires of the different languages have simply been compared, in a diachronic and synchronic way, which makes it possible to see an appearance of the different labels that makes it possible to organize the hierarchy. In other words, it has been possible to confirm that all the languages analyzed have *black* and *white* (or equivalent), most of them also have *red*, and many of them also have *green* and *yellow*. Therefore, with these data, and with a review of the evolution of the languages, it has been possible to confirm that the creation of an independent label to refer to *green* or *yellow* (for colors that were previously included in the *red*, *black* or *white* label) is subsequent to the *red* label. If any attempt has been made to relate salience to the order of the hierarchy, in individual languages, it has been by way of ranking the identification of labels on a continuum of colors in psycholinguistic studies. Although there are some studies of salience from a quantitative perspective in linguistic corpora, they have always been carried out in English or one of the best-known European languages (German, Spanish, Russian, etc.), as we shall see below.

To put it another way, if we identify that a language has the equivalent of *brown*, it means that it will also have: *blue*, *yellow*, *green*, *red*, *white* and *black*. If this same language adds any further distinction to the color continuum in the future, this color will be *purple*, *pink*, *orange* or *gray*. In fact, there may be some dialects of the language or areas where this extension has already begun or where the *brown* label has not yet been added.

When the salience of colors in different languages is studied, it is usually analyzed in eliciting. That is, we look at what position (or how many seconds or milliseconds) the color is identified in a Munsell palette or other similar experiments regarding naming. This would correspond to psycholinguistic experiments. However, we believe that one of the drawbacks of what is known as "work from the lab",[35] typical of the more cognitivist functionalism, is that it does not provide actual frequencies of live language use. Therefore, although they are very valuable data and a very good starting point, linguistics can benefit from real texts with significant frequencies and not eliciting, which is what we propose in this study.

Moreover, this implicational hierarchy has undergone several criticisms from a linguistic point of view and also some modification (as a result of these criticisms), as is the case of the reinterpretation of the simultaneity, or not, of the colors green and yellow in the scale. The most notable criticisms are those made by Levinson[36] and Saunders.[37]

Barbara Saunders[37] reviews from a methodological (study design) point of view the approach of Berlin and Kay.[1] The main problem highlighted by the researcher is the Anglocentrism behind the proposal, which is a common bias in the field of linguistic typology and studies on universals. She comments, for example, that there are some languages that use the criterion of brightness as a distinguishing feature between different color labels. As the English language does not make this distinction, it is ignored as a possible criterion. That is to say, the proposal, due to its universalist aspirations and its bias towards the English language, tends to a simplification that can generate complications when fitting certain languages into the scheme that had not been foreseen (because they are not yet well known).

Saunders' second main criticism of the implicational hierarchy of colors focuses on the source of information, that is, how the data used to create this hierarchy have been collected. The original study is composed of data from 98 different languages. However, only 20 of them have been tested by psycholinguistic experiments using Munsell palettes, as mentioned above. The remaining 78 languages are based on second-hand data, that is, data extracted from what is said in books written by other authors (grammars, ethnographies, word lists, ...) and, obviously, no one can assure us that the 78 researchers have followed the same criteria when collecting these data, a fact that may distort the analysis. Moreover, in the case of the data collected by Munsell, it is not usually seen as too correct that the sample is only 20 speakers for each language and that, moreover, in all cases they are bilingual speakers (with English), a fact that can distort the results. It is also important to note that the minority languages collected in such a study are usually all located in the same geographical area (North America) and therefore there is no genetic and area disparity required to extrapolate typological data as a truly universal trend.

On the other hand, we also have more concrete examples of specific languages that are exceptions to the color hierarchy, as presented by Levinson.[36] Leaving aside the fact that there are some languages in which there is no

word for the meaning of 'color', there are Papua New Guinean languages that compose their colors from reduplications of other objects or entities, as is the case with the color *black* which uses the reduplication of the word *night*. There are also a large number of languages that designate the color *orange* as fruit, something that occurs in English itself. This would be in contradiction with another of the criteria Berlin and Kay present when designing their hierarchy, since they comment that colors should be abstract and not designate concrete objects in the world.

Recently, within computational linguistics, there have also been analyses that try to revise the proposal with different techniques that allow us to obtain new data. McCarthy et al.[38] want to confirm whether the criteria of abstraction, monlexical forms and salience, proposed by Berlin and Kay, are met computationally. While this study provides significant data, it does not delve as deeply as we might hope into the topic of salience, which is of particular interest to us. These authors mention that there are not enough linguistic corpora to carry out an impeccable study from the point of view of typological balances (true fact) and, for this reason, they comment that they will not analyze the corpora computationally and will simply review the data already provided by Berlin and Kay in the World Color Survey,[6,39] that is, eliciting data from psycholinguistic studies.

As a consequence of all these reviews by different authors, the original color hierarchy proposal has been slightly improved (or extended) in recent publications.[6,39–41] In general, it can be noted that more languages have been analyzed. These new data have been collected through a unified methodology and using exclusively Munsell (not second-hand data from grammars), in the World Color Survey. The selected languages are more closely related to balances and typological criteria.

As a consequence of these improvements and the power of the proposed hierarchy, we must affirm that it is still a very accepted and widely used theory for research related to colors in languages. For this reason, in recent years, we can see studies on the functioning or application of this hierarchy in a specific language such as Czech[4] or Hungarian[5] or in general in linguistic families such as the Slavic languages.[42] There are also authors who have opted to extend the methodology proposed in the World Color Survey to languages that had not been included in that study,[7] with a dialectal detailing with differences in the coding of the basic colors.

Other more recent studies, applied to languages such as Italian[43] or Russian,[44] try to offer new data in these languages in order to be able to specify and improve the theory. In the case of these languages, moreover, there is a latent interest in resolving the existence of two basic colors for what in English would be collected with a single term (*blue*) in the hierarchy. In the case of Russian, we can also find examples of diachronic studies for other specific terms in the basic color hierarchy such as *brown*, with the analysis of the competition of different labels over time and their monitoring thanks to corpus analysis based on literature.[3] In the aspect of diachrony it is very interesting to highlight the work on the Nafaanra language[45] which shows a clear dynamic application of Berlin and Kay's hierarchy in stages (within a difference of a few decades). That is, Zaslavsky et al. show how in the late 1970s, in that language, only stages 1 and 2 of the hierarchy could be attested (up to the color *red*). Today, however, four decades later, 10 colors of the scale can be attested. This rapid change is probably due to the greater penetration of English (thanks to the globalized world we live in) in that language and to the current designational needs in that culture (also a product of recent changes).

In relation to the studies of the link between basic color terms, salience and quantification (in corpora), we must highlight the proposals of Steinvall,[2] Corbett and Davies,[46] Hays et al.,[47] Johansson and Hofland[48] and McManus.[49,50] In general, we must say that these are proposals focused on the English language (or some other European language) and of very specific and limited corpora. Therefore, our proposal aims to offer more varied data, drawn from many more languages, in order to make its application more universal.

2 | METHODS

2.1 | Sampling

In recent years there have been several proposals for labelling linguistic corpora. Some of them, moreover, have attempted to do so with the same system and the same tags for all languages, which would allow the results to be compared. Probably the most famous proposal is Universal Dependencies,[51] which contains corpora annotated by dependencies and with a very clear intention to be universal (i.e., to be used for linguistic typology). Its use is especially intended for syntax, although its labelling of words in lemmas allows for a good study from a lexical point of view, which is what we are concerned with here regarding colors. Thus, our study has been carried out specifically thanks to Kontext.[52] This resource has in fact been designed to process raw data from Universal Dependencies, and it is integrated within the multiple tools developed under Lindat-CLARIN. There are two main reasons for this choice. Firstly, is that Kontext has a very intuitive and visual interface. In other words, it is user friendly. Secondly, Kontext not only

contains the Universal Dependencies corpora (so the data are the same, but with a more practical appearance), but also other corpora annotated in a homogeneous way by others, such as Parseme or the European Union (EUDGT), which are also parallel.

As is well known, it is impossible to review the data for all the world's languages (or even to know how many there are).[15,53,54] There are many reasons for this: there are missing data for many non-European languages for which we only have the name,[17,19,55] there are languages new languages and languages that die,[14] or we have not yet agreed on the distinction between language and dialect.[56]

Therefore, we must create a representative sample of the world's languages to be able to extrapolate the results as universal.[57] There are three different types of linguistic sampling[58]:

- Variety Sampling. If the phenomenon is poorly studied and we want to know its full development, linguists often opt for a variety sampling, which is a selection of representative languages, including languages that may provide unusual examples.
- Probability Sampling. If we want to see the percentage of universality of a phenomenon, we opt for probability sampling, which is a selection of languages taking into account balances of representation of different language families, areas and linguistic types.
- Convenience Sampling. If the availability of the data is not rich enough, the sample has to be created without taking into account the perfect equilibrium described above.

What we propose for this initiatory study is a convenience sample. In the corpus-based research we want to carry out, it is not possible to propose a typologically perfect selection of languages as discussed above. The problem lies in the few languages that have homogeneously labeled corpora valid for such a study. This selection is characterized by the use of all the resources available to the researcher, without taking into account these trade-offs. This often occurs in the first studies, as was the case with Greenberg's selection of 30 languages in his foundational proposal[10] or with the same Basic Color Terms hierarchy.[1] The main objective is to present results that are acceptable and indicative of a trend which, later, when much more data are available that allow for greater balance, will be profiled or confirmed. In other words, all the available languages of the resource to be worked with will be analyzed. However, the characteristics (typological, genetic, areal) of the different languages worked on will be taken into account while analyzing them in order to try to mitigate as far as possible the possible biases that occur in linguistic studies of this type, produced by what may be called WEIRD languages[59] (from Western Educated Industrialized Rich and Democratic countries).

The last problem to be highlighted is the fact that we find corpora in some languages, but with an insufficient number of tokens for the study we want to carry out. By token, in corpus linguistics, we mean the total number of words that appear in a corpus. If the same word appears twice in a corpus, for example, it will be counted twice for the purposes of tokenisation. In addition, other elements such as symbols or digits are also considered tokens. In other words, an excessively small corpus will not be able to display enough instances of each color to allow for a reliable comparison. After an analysis of the behavior of the different corpora, it has been found that we cannot have guarantees with corpora of less than 40 000 tokens. However, it is recommended that the corpus should have more than 100 000 tokens, which allows a much better comparison. We also believe that these figures and criteria are not only of interest for the study of colors, but are proposals that can be extended to other lexical and semantic aspects. Languages with a very small number of tokens have been discarded from the corpus. Languages with a number of tokens close to 40 000 tokens have been analyzed and the data can be presented with or without these languages. In other words, this influence can also be assessed.

The various treebanks have also been reviewed, and those languages in which errors in the search made it impossible to obtain conclusive results have been discarded. Therefore, in total, 57 different languages will be analyzed, including languages whose number of tokens is borderline of the set limit. In the case of data above the threshold, the analyze will be done in 46 languages. As will be seen, languages that can be considered dead have not been discarded, as they can provide us with valuable information from a diachronic point of view and their relation to the implicational hierarchy and the different stages. Table A1 (in the Appendix A) shows these analyzed languages, where the 11 languages that are not analyzed in the 46-language version are marked in orange. We also indicate the number of tokens we will analyze from the different languages, in relation to availability and whether it is a language belonging to the Indo-European family or not, to highlight the languages that are less studied. In other words, as our main interest is to quantitatively document as many languages as possible from corpora, we take into account all the languages that are available in the mix of resources we have used. Therefore, 57 is the maximum number of languages we can work with and will therefore be the number of languages we work with.

Another aspect to be decided before starting the research in order to extract specific data is the type of corpus to be analyzed. It is well known in corpus linguistics[60]

that depending on the type of analysis to be done, the texts on which the corpus is based are required to be of a certain type: fiction, non-fiction, legal, wikis, forums, news, social networks, etc.

We have decided, for our study, not to make any kind of selection and to take into account all the varieties of corpora available. The justification is twofold, again. Firstly, as we have already mentioned, this is an innovative study and, therefore, the aim is to collect as much data as possible. If, for example, some researcher, in the future, wants to carry out a study based on our data, but taking into account only one corpus typology, this would be possible. Secondly, if we take into account different corpus typologies, we will have a more varied and realistic picture of texts from different communicative situations and social groups, which will ensure greater confidence in the extent of color usage equally throughout the speech community. We do not present all these data in this section because they take up too much space. The repository linked at the end of this paper includes the specific data pertaining to each corpus, in each language. Therefore, it is possible to play with different variables and, in the end, we will analyze 136 different corpora in the 57 languages announced.

2.2 | Data

To review the behavior and appearance of the different labels referring to colors in different languages, we must ensure that we have a unified and secure way of knowing what the colors are called in those languages. The difficulty is clear: some of these languages are little known to researchers in particular and to the world in general, and it is necessary to have a methodology that offers certainty in order to obtain the labels, which in most cases will be unknown. In addition to sharing a single homogenized criterion, the tool or source must allow for possible synonyms or alternative versions of mentioning a color for what would correspond to a single color in English. It is also important to obtain the words as a lemma (i.e., the word without adding any modification).

The most common approach so far has always been to go directly to the particular grammars of each language (or similar lexical lists and second-hand documents). The problems of following this widespread methodology are basically twofold. Firstly, each source for each language has been elaborated in a different way, there is no unified methodology. Therefore, depending even on the author responsible for the source, different criteria may have been used to collect this information. Secondly, the data may not be up to date. On the one hand, the form of words may be slightly modified over time, which would not allow us to find such a word in current corpora (since many language grammars with little data are decades old). On the other hand, we should remember the problem reported by Zaslavsky et al.[45] in the case of Nafaanra: in 40 years the language has greatly expanded its repertoire of linguistic color distinction. Therefore, if we were to go to a grammar from the 1980s, we would find 3 labels instead of the current 10 (which are the ones that could potentially be in the texts we want to analyze).

For these reasons, in our research we have worked with the tool already used in a study on colors by McCarthy et al.[38]: PanLex.[61] In this database we have up-to-date data for virtually all languages that have documentation. The words appear as lemmas, which makes it easier to search for all possible forms of the word. It also clearly shows the proximity of meaning between two lemmas, that is, possible synonyms. We have set the criterion of considering as synonyms those lemmas that share more than 66% similarity. Subsequently, we have confirmed that it is a synonym through the process of double translation with English. For practical purposes, we have also considered as synonyms different ways of referring to a color that are dependent on their syntactic function (as may occur in languages such as Japanese). With the double translation process we have also been able to check whether there are any cases of homonymy (words with different meanings, but spelled the same) in the languages. In the case of detecting homonyms, such as *horia* in Basque, which refers to the color *yellow*, but is also other things (such as a relative like *that* in English), this has been marked for later solution when working with the corpora. In short, by using this tool and following these steps, we aim to reduce the anglocentric bias of such a search.

The data relating to the 57 languages we have analyzed, with all the Basic Color Terms and the different synonyms and other versions that have been taken into account, are freely available for use by any researcher in a repository, as detailed at the end of this paper, contacting to the corresponding author.

2.3 | Calculating salience

The raw data we have obtained from the analysis of the different corpora are not comparable, as there are some with millions of tokens and others with a few thousand. As a result, we have chosen to work on the Index per Million Tokens (IPM). In this way, all color occurrences are made on a same common universe and the results are comparable (in general or between languages, corpus typologies, etc.).

In addition to the raw number of occurrences and the number per million tokens, we also offer the Average Reduced Frequency (ARF), in Equation (1), which is the application of a result correction formula that guarantees homogeneity in the distribution of the study element. In other words, the result obtained after applying this formula guarantees that the appearance of this color is not concentrated only in one text of the corpus or in a very specific part of it, but that it is distributed in a balanced way throughout the different texts that make up the corpus.

$$ARF = \frac{1}{v}\sum_{i=1}^{f} min(d_i, v) \quad (1)$$

where f is the frequency of the given expression in a corpus of the size N, d_i is the distance between the individual occurrences of this expression in the corpus, v is the average distance between its occurrences.

3 | RESULTS

After carrying out our analysis, we have obtained results that match very well with Berlin and Kay's hierarchy of stages, as shown in Table A2 (in the Appendix A) and Figure 2. The results to be discussed will always be based on the IPM format. The data we offer in both outputs as "full" correspond to the totality of the languages available in the resource used and which we have been able to analyze (all 57 languages). However, we also offer a slightly smaller sample, named "clean", which corresponds to the 46 languages that have a corpus with a size greater than 40 K tokens and which, therefore, yields more extensible results. In Figure 2 we have added a red line that allows us to identify more clearly the two groups that we believe constitute Berlin and Kay's implicational hierarchy and which we further develop in the discussion.

In relation to the universal and original distinction between *black* and *white*, we must say that there is an overall higher frequency of *black* than *white*. This fact contrasts clearly with the data provided by the researchers of individual European language corpora, who point out that *white* is more frequent,[47] except for Steinvall,[2] who shows in the Bank of English corpus exactly the same distribution as we do. Therefore, we have decided to check in how many of the languages in the corpus *white* or *black* appear in the first position, illustrated in Figure 3. We should also mention that, if both colors have the same weight, that language has not been counted in Figure 3.

In relation to the second group proposed by Berlin and Kay, consisting of *green* and *yellow*, the preferences are much clearer. While both may correspond to the same evolutionary stage, *green* is more frequently used and easier to identify as shown in Figure 4.

This situation, which is completely different from the situation with *black* and *white*, can also be seen in the frequencies of the individual languages. For this purpose, we have elaborated a graph for *black* and *white* (Figure 5) and another one for *green* and *yellow* (Figure 6).

The last group of colors proposed by Berlin and Kay is the one related with the last evolutionary stage, the least basic colors within the basic colors. As can be seen in the overall graph, all of them have a much lower frequency, although there are many more occurrences of *gray*. To confirm this trend of total frequencies in million tokens, we have also gone to look, individually, if the

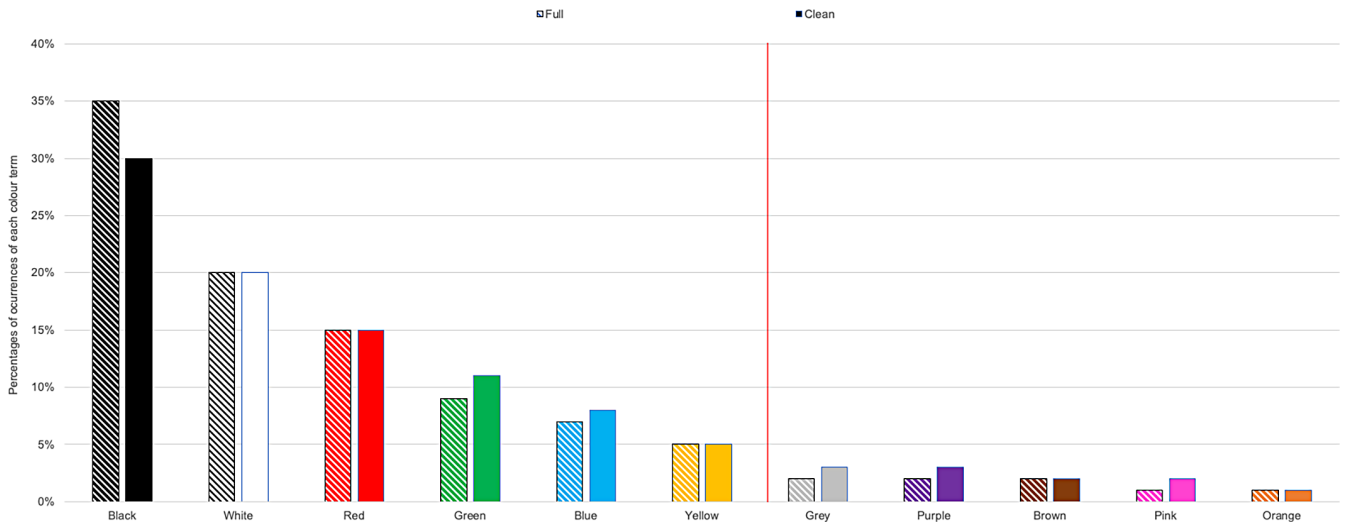

FIGURE 2 Hierarchy based on frequencies - percentages on IPM.

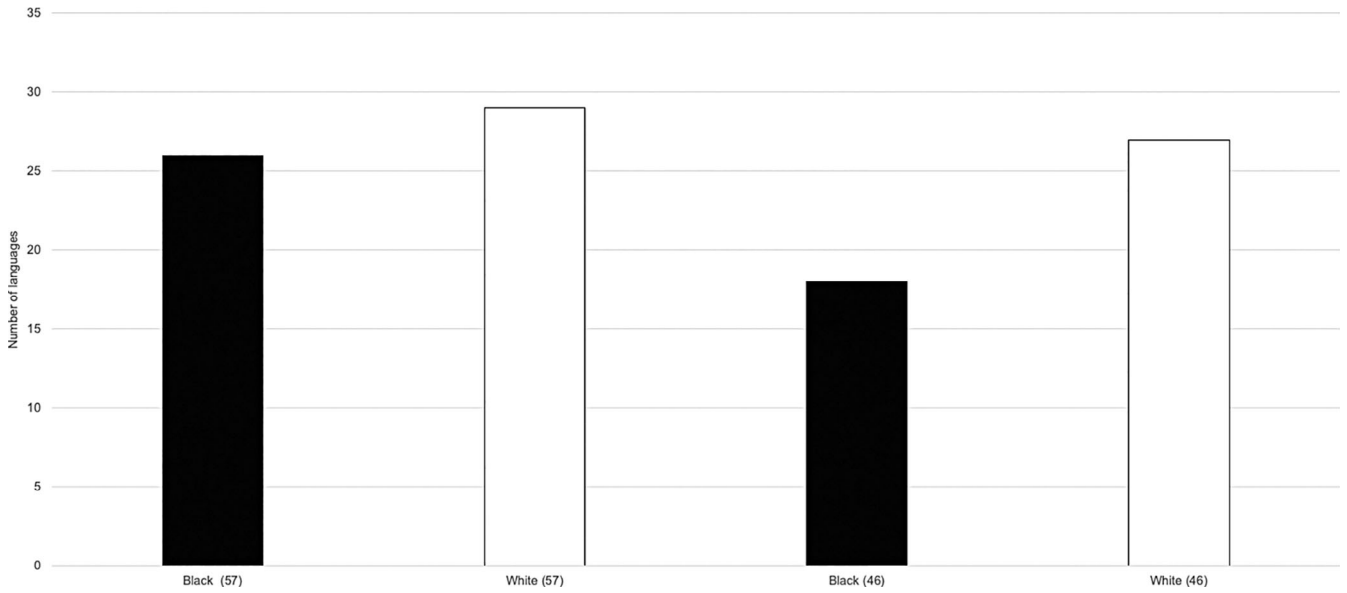

FIGURE 3 Dominance of *black* or *white* in our corpora.

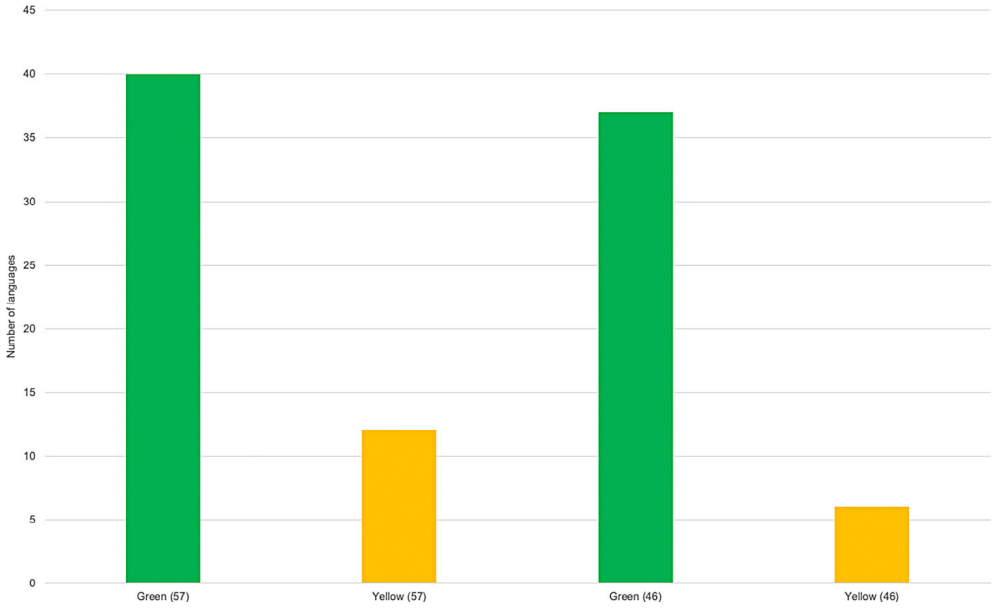

FIGURE 4 Dominance of *green* or *yellow* in our corpora.

majority of languages choose to highlight gray ahead of the rest, as shown in Figure 7.

4 | DISCUSSION

As is usual in typology studies, when studying languages of such diverse types, problems or difficulties arise that were not foreseen, because these distinctions are not relevant in English or similar languages. For example, it is important to make sure that the search is done using lemmas and not words. Therefore, since we have the lemmas available through PanLex, we can search directly for all the different morphological forms of the same color depending on the context in the sentence. In other words, not only we will get the colors in their plural, feminine, etc. versions (if that language has such distinctions), but also all the declined forms in languages with many endings such as Polish or Russian. Thus, we are not only taking into account *czarny* in Polish (*black* in its nominative masculine singular form), but we are also taking into account: *czarne*, *czarnego*, *czarnym*, *czarnej*, *czarna*, *czarnych*, etc.

The main problem we have noticed when carrying out our research is that there are languages that change their words even more depending on the grammatical

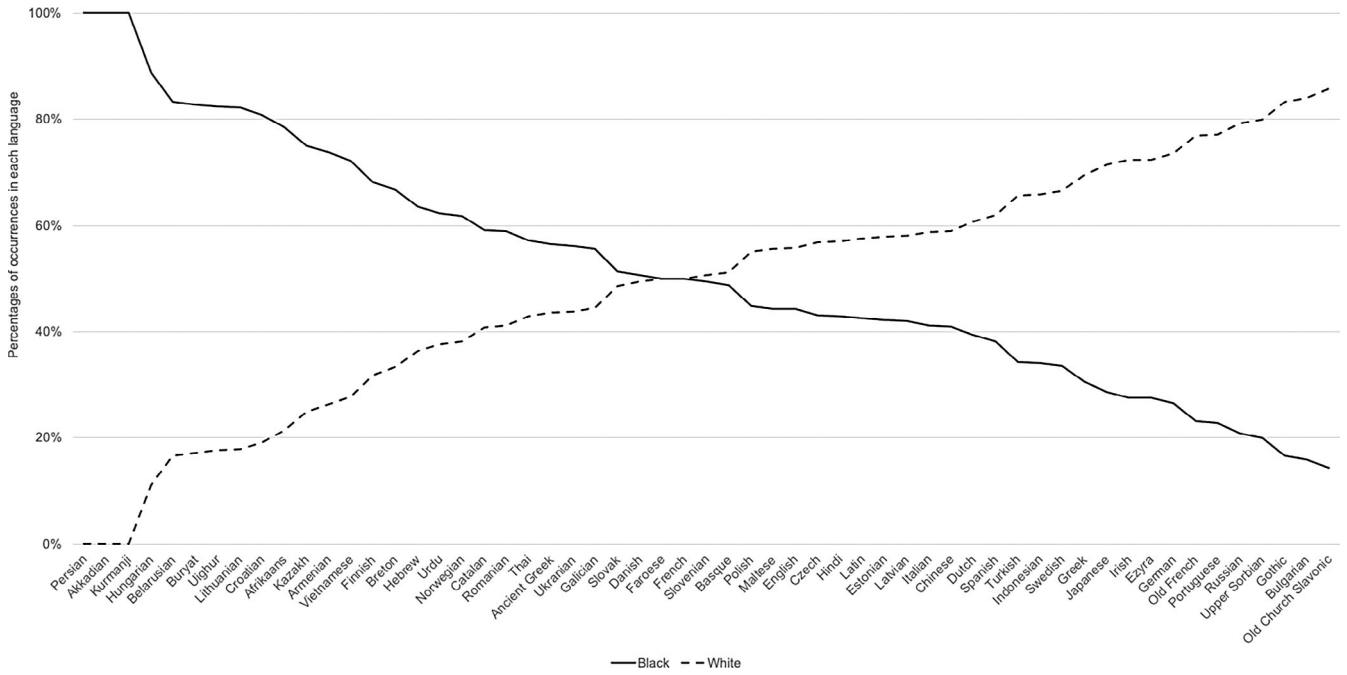

FIGURE 5 Percentage of *black* and *white* in each analyzed language.

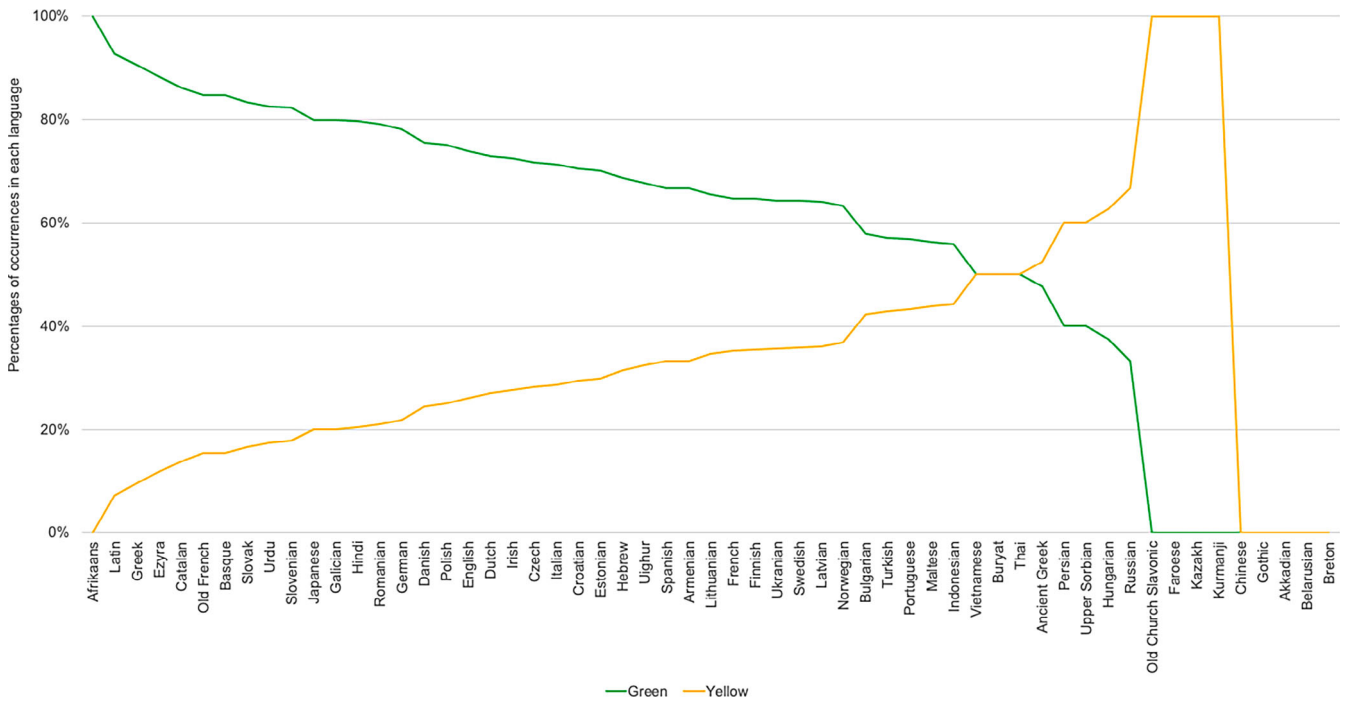

FIGURE 6 Percentage of *green* and *yellow* in each analyzed language.

category or syntactic function of the word, the root changes completely, as in the case of Japanese. In these cases, we have chosen to include all the different versions of the same color detected in PanLex and treat them as if they were synonyms. In this way, we can also see what distribution there is between them and, finally, we can also obtain the sum total of all the versions.

Other languages also make a distinction between upper and lower case. In many corpora such a distinction is established and our interest is to collect all the versions. For this reason, we have configured Kontext

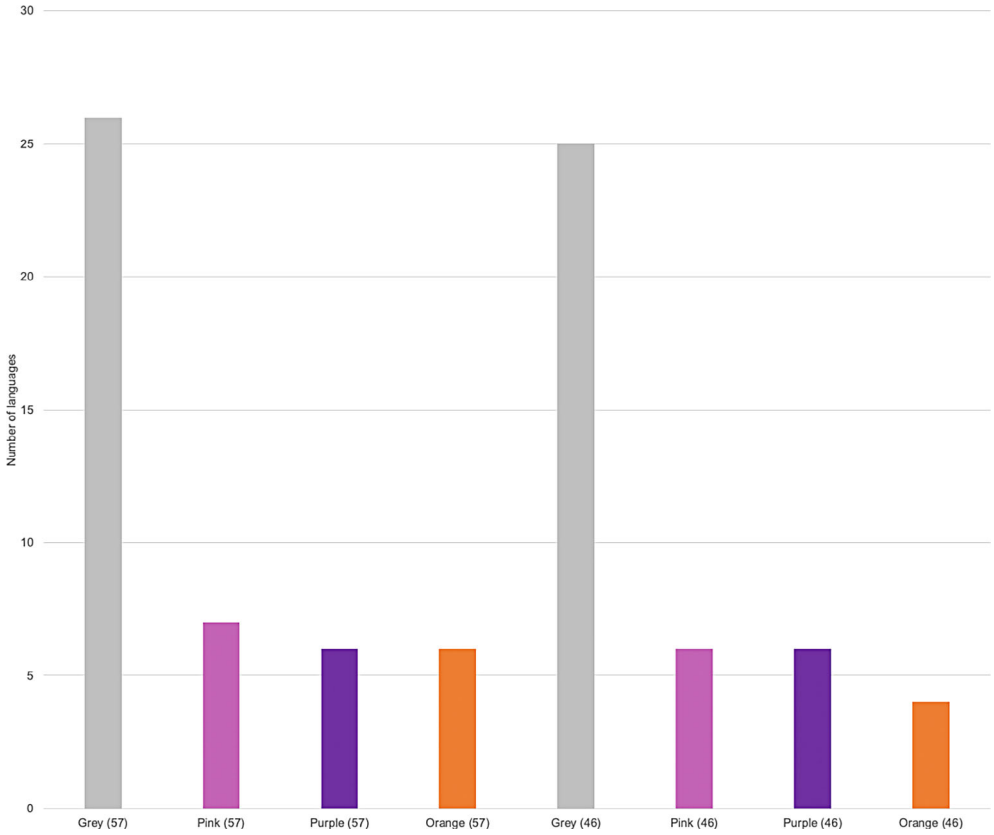

FIGURE 7 Dominance of the last 4 colors in our corpora.

to be case-insensitive in order to obtain all mentions of a given color.

The last difficulty that can occur in languages is homonymy, as mentioned earlier. If this problem is not solved, the numbers of occurrences of a color would be abnormally high. Therefore, we have applied a QCL (query) in Kontext to all cases of homonymy that we have previously detected with PanLex and the double translation. There are two possibilities to configure the query. On the one hand, we can take into account only adjectives, as, for example, in: $[upos = "ADJ" \& lemma = "azul"]$. In case we prefer to include all grammatical categories except the one that generates the case of homonymy, we simply have to include in "upos" all possible categories except the one that generates a conflict.

Surely, when more languages are available, it will make more sense to use the second option, as there are some languages (in very rare cases) in which colors may not manifest themselves through the adjective category but through other more specific categories. In spite of this, whichever option is chosen, logically, we must be consistent in the analysis of all corpora and always use the same choice, as well as explicitly mentioning it and the implications it entails.

It should also be mentioned that due to space and time limitations we have left the colors outside of Berlin and Kay's original hierarchy without a detailed exploration (which we wish to carry out in future work). It would be interesting, therefore, to know in detail also the frequencies of these non-BCTs colors (according to Berlin and Kay) and to relate them to the results obtained in the canonical labels of the classical hierarchy.

The first aspect we can comment on Figure 2 and Table A2 (in Appendix A) is the proportionality that exists between the analysis of 57 languages (which includes some corpora with a low number of tokens) and 46 languages (which contains languages with more than 40 K tokens). Generally, we can say that the distortion of including such low corpora is not so large and that the results are still valid. The biggest distortion occurs at the extremes of the scale. On the one hand, an abnormally high number of *black* color appears, corresponding to a (not necessarily high) occurrence in very small corpora which, when adapting the result to 1 million tokens, grows exponentially. On the other hand, there is an abnormally low number of the last four colors (the last stage in Berlin and Kay), since in such small corpora there is no margin of occurrence of these colors.

Also relevant is the internal distinction that can be made in the scale between the first 6 colors and the last 5, marked in Figure 2 with a vertical red line. In fact, from frequency, the color *brown* could be grouped with the last 4 colors, as it behaves in a very similar way. The frequency between them is not decreasing and is equally

low, never exceeding 20 occurrences per million tokens (approximately). Although Berlin and Kay wanted to give a special entity to *brown*, highlighting it before the last group, we want to emphasize that its salience is at the same level as the rest of the colors mentioned (*pink*, *gray*, *purple* and *orange*). Although *brown* was temporarily created before these colors, its use is not as established as *blue* or *yellow*, for example.

Another striking aspect of the results obtained is the decreasing and quite proportional order that can be observed in the graph. This drawing helps us to believe that there is a clear and direct link between salience and the stage proposed by Berlin and Kay. Therefore, we join the proposal of Corbett and Davies[46] and McManus[49,50] who report a high correlation between frequency and the evolutionary hierarchy of Berlin and Kay (tau = 0.77, $p = 0.001$) in their studies in English or other big European languages. From an evolutionary-functionalist perspective, the results make a lot of sense: the older the color created, the more frequent it will be, as it will be more deeply rooted in the language and the minds of the speakers who use it. The later distinctions (the late-stage colors) are supposed to serve to much smaller designation needs, which are consequently less present in texts. Therefore, the initial hypothesis is confirmed. Moreover, these data invite us to further research work, as we believe that behind such a data arrangement is Zipf's Law, because of the shape of the layout of the results that we see in the Figure 2.

As shown in Figure 2, the only color that breaks the decreasing order is *yellow*. Far from being an imperfection of the analysis carried out, this result was fully expected after a thorough review of the literature and confirms the proposed study model. If this apparent exception appears in *yellow*, despite being less frequent than the following color, *blue*, it confirms the theory proposed by some authors such as Steinvall,[2] who even dedicates a section to "the problem of yellow". In his study, in line with the research carried out by Corbett and Davies[46] and McManus[50] in English or Hays et al.[47] in French, Russian or German, *yellow* shows a lower frequency than blue. The explanation offered by these different researchers is that there is a relatively important weight of colors not considered basic (according to the previously exposed criteria) such as *gold*, *Golden* and *blond(e)* that diminishes the occurrences of *yellow*. In other words, this low frequency of *yellow* according to its position on the scale is due to the fact that this color is not structured in the same way as the rest of the other categories of primary colors.[2] In short, this tendency, which has been pointed out by different researchers in few European languages, can be extended to many other languages, Indo-European or not, thanks to our data. Therefore, after seeing this tendency in practically all 57 languages analyzed, this specificity is closer to being a candidate for a universal pattern. There is also the possibility that the salience of *yellow* is just lower than we thought because of the location in the Berlin and Kay proposal. However, such a strong assertion cannot be made without more data pointing in that direction.

One of the main benefits of the results presented in this overview graph is the confirmation or dismissal of some of the ideas previously formulated by other researchers on specific languages. It is worth remembering that, in addition to the fact that there are only a few languages analyzed with corpora in the literature, these languages have been studied using different, non-unified methodologies and it is therefore much more difficult to compare the data. In the case of our analysis, the unified data allow comparison and presentation of a general classification, as well as the study of languages never before reviewed through corpora quantitatively, as Abdramanova[62] points out regarding Kazakh, a language we analyzed: "to sum up, in Kazakh linguistics, to notion of colour basicness is not clearly defined by researchers; it could be assumed that the conclusions elicited from their studies are based on two main grounds: (a) they are psychologically highly salient and (b) they are most frequently used colour denominations (though no frequency data have been presented) in the language of fiction, media, and oral discourse". Specific data relating to these languages, with or without previous studies, are also available in the repository we use to locate our data.

On the other hand, we should mention the findings obtained in the comparison of individual language-specific data and similar information found in the literature, using other sources or other methods, as we showed in Table A3 in the Appendix A.

Similarly, our data also allow a comparison with some of the few languages that have been analyzed from frequencies. In this way, we can enrich and amplify the knowledge we have on these languages. For example, we can compare the results obtained in this paper with our method for German with those proposed by Hays et al.[47] in this language, as showed in Table A3, in the Appendix A, where the ranking of the colors in the two works is presented.

Firstly, it should be noted that the order of the colors of the German language is roughly the same in our proposal as in that of Hays et al. This fact allows us to reinforce some of the results proposed by these researchers which are not frequent in other languages, but do occur in German. Or, in the same way, we find results that validate and reinforce our research. One aspect of this is the order of the first three colors: *white*, *red*, *black*. In the German language, *white* is more frequent than *black*,

and *red* is also more frequent than *black*. This is not only the case in literary corpora, as in the case of Hays et al., but also in administrative corpora, wikis, news, technical, blogs, etc.

Secondly, we must show how our analysis can add a greater degree of precision to the proposal of Hays et al.[47] In this case, these researchers do not offer any frequency of occurrence for *pink*, *purple* and *orange*, thus relegating them to the last level, without any possible distinction. In our case, we can also propose a hierarchy for these three colors: the most frequent of them would be *pink*, then *orange* and the least frequent by far would be *purple*.

Regarding data from Figure 3, the distinction between *black* and *white*, as can be seen, in the case of the analysis of 57 languages, the distribution is practically homogeneous. That is, there is no striking preference for one of the two colors. In the analysis of 46 languages, we can see that there are 27 languages in which *white* predominates and 18 languages in which *black* is the first choice. These data, together with the above-mentioned high frequency of occurrence of *black* in languages with a corpus of few tokens, lead us to believe that our study is uncovering a possible bias of the studies done so far, based on WEIRD (Indo-European) languages. We believe that, once languages with fewer tokens have more annotation, the number of frequencies per million of *black* will decrease. However, the tendency to give more prominence to *black* over *white* will persist. Just as the dominance of *white* will prevail in the languages already analyzed. In short, we believe that the best approach in this case is to return to Berlin and Kay's original proposal,[1] in which they do not establish any kind of distinction between the two colors. That is, we believe that no color is more frequent than the other in a general way, since the existence of one entails the existence of the other, as is the case with the verbs *buying* and *selling* or with *yin and yang* (which is why, probably, these colors are used for the representation of them), they are mutually dependent. Therefore, the preference for *white* in the languages analyzed by other authors (German, Spanish, French), for example, may be due to area, Indo-European, cultural or extra-linguistic influences, or they may simply be coincidences of the corpus analyzed, but in no case constitute a universal tendency.

As can be seen in Figure 4, 40 languages show a strong preference for *green*, while 12 show a strong preference for *yellow*. In the rest of the cases up to the 57 analyzed (5 languages), either the frequency is the same or they do not have either of these two colors in the full analysis. And 37 prefer *green* and 6 *yellow* in the reduced analysis. Therefore, we believe that in this case we can confirm that *green* is, in general, more salient (in frequencies) than *yellow*. In individual language studies from a psycholinguistic perspective we do not find agreement and each study shows different options, such as the preference for *green* in Kazakh in Abdramanova[62] or the preference for *yellow* in Czech in Uusküla,[4] for example.

The difference between Figures 5 and 6 is clear. In the case of *black* and *white* it can be seen how the path of both colors is quite similar, even parallel in some cases. It should also be noted that the change in dominance from *black* to *white* occurs in the center of Figure 5. This assumes that there is no clear constraint on such dominance. Therefore, as we have indicated, the preference for one or the other is rather anecdotal and that is why we speak of the equality of *black* and *white*. On the other hand, in the case of *green* and *yellow*, we can see how in practically the entire graph *green* is in a superior or equal position to *yellow*. In other words, the shift to *yellow* dominance over *green* occurs towards the end of the Figure. Therefore, we can affirm that the greater salience of *green* (in terms of quantity) is regular in very different languages and preferred by them.

Regarding the last group of colors, showed in Figure 7, both in the analysis of all languages, and in the analysis that does not take into account some languages with few tokens, *gray* is the most frequent choice. Therefore, it is not only just greater occurrence in general, but also of greater occurrence within each language, with a few exceptions. In fact, in some corpus studies such as Steinvall's,[2] its frequency is higher too. In psycholinguistic eliciting studies we can assure the same, as in Abdramanova's,[62] for example. Therefore, being consistent with the hypothesis of our study, we believe that the first of the distinctions established in this group was probably *gray*, whose appearance is more frequent than that of the other three colors, which probably appeared later in the rest of the languages.

It should also be noted that in the general hierarchy, in the case of *brown*, *gray* and *purple* the occurrences observed are very similar. The distance is so short and the values are so low that the fact of showing one of them before the other is mainly due to the fact that, according to our data, this is how they should be placed in the case of a relative tie. We only want to point out that the important thing is to isolate the 5 final colors from the rest of the colors and that these 3 colors have a superior entity compared to the other 2.

In relation to the last of our objectives, we wanted to review different forms for a single English word. All the data are collected in full in the repository. What we would like to stress is that in this aspect we are not so interested in the particularity of the data, as these can still be controversial or debatable, as there is no consensus in many of the cases. In addition to the lack of unanimity and the fact that a high level of linguistic knowledge is required, in many languages we still do not have enough data to know the precise situation or relationship of two or three particular words. For this, we believe that a quantitative review

of the occurrences of different words with the same meaning can help to better understand the strict relationship between them. For example, in the case of Slovenian, for *yellow*, PanLex gives us two different words *rumen* and *žolt*. The difference in frequency of occurrence is quite remarkable: 22 *rumen* versus 4 *žolt*. Therefore, we can propose that *rumen* is the more basic, current, unmarked or frequent in Slovenian. Obviously, if we do not have language-specific information, we cannot know to which specific criterion such a difference is due, but this is precisely an aid to find it out and to predict these aspects. In this case, as we have enough information on the Slovenian language, we can find out that what is happening in this case is related to diachrony: *žolt* is an older version and is therefore falling into disuse. On the other hand, *rumen* is gaining ground.

There are other cases, such as in the differentiation of *gray* in Lithuanian, where the difference is very low: *pilkas* (1.52) versus *širmas* (0.09). When more data become available in the future, we will be able to confirm the trend that we can point to at present: *pilkas* is more basic than *širmas*, either because it is less contextually bounded, or because of greater use by the general population, or because of extralinguistic issues. Surely, *pilkas* could be considered the hyperonym and *širmas* the hyponym, that is, the less frequent word has more semantic nuances (in this case being a 'lighter gray') and hence its lower frequency.

The last controversial aspect we would like to highlight is the difference between *purple* and *violet*. From an optical or physical point of view, it is much easier to consider that they are two different colors, after reviewing their technical properties. From a linguistic point of view, it is more complex to determine whether they can be considered synonyms or not. The main problem we find is, on the one hand, the vagueness in the concept of synonymy (as it is gradual, not discrete). On the other hand, the use of language does not necessarily have to be closely related to reality. In fact, this aspect is often part of the idiolect of each speaker and, therefore, different opinions can be found in speakers of the same language.

When looking at occurrences in Polish, for example, we find that *purpurowy* appears 0.10 times per million words and *fioletowy* 9.73 times. This disparity of occurrence could be used to argue that, although they are different colors in reality, these labels act as synonyms in the Polish language. We understand that there is no such precise designating need in our reality that Polish speakers are forced to mention the label *violet* 10 times out of a million words, while *purple* is hardly ever mentioned. In other words, we do not believe that there are more objects in reality that need to be called *violet* than *purple* and, therefore, we believe that it is a choice of the speakers. Thus, we believe that even when a speaker is faced with a purple object, he or she chooses to ignore the brief nuances that characterize it as *purple* and names it as *violet*, a more common label in Polish. In short, following this hypothesis, *violet* is a more basic form of this color in Polish, for example, and *purple* would act as a more marked form. In other words, *fioletowy* would be a more general word, applicable to a wider range of the spectrum, while *purpurowy* will be used only in very specific contexts, where a precise distinction needs to be made, or by very specific speakers who always choose to make this distinction. That is, de facto, *purpurowy* would be functioning as a "type of" *fioletowy*.

5 | CONCLUSIONS

This paper presents new quantitative data on the salience of the basic color terms proposed by Berlin and Kay,[1] based on real texts from corpora, which represents one of the main differences from previous approaches. Our results were obtained from the analysis of 57 different languages and 136 different corpora. As our data represent a greater variety of languages compared to previous studies,[50] we have provided hierarchies based on new frequencies.

Salience is a criterion that can also be used to order the basic colors of a language (and not only to differentiate between basic and non-basic colors). Our results show that there is a close relationship between the implicational hierarchy and the frequencies of its salience. Our data confirms Berlin and Kay's original proposal, except for *yellow*, by showing a quantitatively proportional decreasing order. The way the frequencies decrease is quite exponential and refers to Zipf's Law. These results (from corpus and in 57 different languages) are aligned with previous salience studies on basic color terms.[2]

These data also agree with other proposals of other authors based on other methodologies (experiments) and in one language (English), as in Reference 63. Therefore, by means of this confirmation through different ways we may be unveiling a clear restriction, what the previously cited authors group, elegantly, as: primary colors (the first six colors) and basic colors (the primaries + the other 5). It is also interesting the coincidence that can also be traced between the group they call achromatic (*white*, *black*, *gray*) with our finding that, of the group of non-primary basic colors, it is precisely *gray* that stands out.

A review of the correlation of frequencies between the colors found in the same evolutionary group formulated by Berlin and Kay[1] (and, therefore, without distinction) is presented:

1. A first group includes colors with high frequencies and a decreasing progression in relation to evolutionary stage: black (35%) > white (20%) > red (15%) > green (9%) > blue (7%) > yellow (5%).

2. A second group includes colors with low frequencies of occurrence: > gray (2%) > purple (2%) > brown (2%) > pink (1%) > orange (1%).

One of the main challenges in establishing the above hierarchy when working with languages different from English is dealing with different synonyms of the same English Basic Color Term. In order to avoid the anglocentrism bias we have considered different term possibilities that the languages included in our set use to refer to the basic colors.

In fact, a further research line could be to use our synonyms from the dataset for the same basic color terms in languages different from English. The results of this quantitative method could be enriched with in-depth qualitative data (diachronic, context-specific, diaphasic, diastratic, dialectal, etc.).

This paper is a first step for establishing a new approach to the basic color terms. The pillars of this new method are the cross-linguistic perspective and the corpus-based approximation. In the future, a greater availability of corpora from different and varied languages will improve our knowledge of colors' salience by avoiding WEIRD languages bias. Although the quantitative approach have limitations, we claim it is useful to explore this methods to complement previous approaches.

AUTHOR CONTRIBUTIONS

Both authors have contributed equally.

ACKNOWLEDGMENTS

We would like to thank the Martí i Franquès Research Fellowship Programme and the Diputació de Tarragona. We would also like to thank the reviewers for their helpful comments in their review.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available for any researcher on request from the corresponding author. All the raw data generated by this research are openly available online in the repository CORA: https://doi.org/10.34810/data773

ORCID

Antoni Brosa-Rodríguez https://orcid.org/0000-0002-8474-2065

M. Dolores Jiménez-López https://orcid.org/0000-0001-5544-3210

REFERENCES

[1] Berlin B, Kay P. *Universality and Evolution of Basic Color Terms*. University of California Press; 1969.

[2] Steinvall A. English Colour Terms in Context. 2002.

[3] Bochkarev VV, Shevlyakova AV, Paramei GV, Rakhilina EV. A quantitative study of Russian colour terms buryj and koričnevyj in the Google books Ngram corpus. *CEUR Workshop Proc*. 2020;2852:1-10.

[4] Uusküla M. The basic colour terms of Czech. *Trames*. 2008; 12(1):3-28.

[5] Uusküla M, Sutrop U. The puzzle of two terms for red in Hungarian. In: Wohlgemuth J, Cysouw M, eds. *Rara Rarissima*. De Gruyer Mouton; 2010:359-376.

[6] Cook RS, Kay P, Regier T. The world color survey database. In: Cohen H, Lefebvre C, eds. *Handbook of Categorization in Cognitive Science*. Elsevier Science Ltd; 2005:223-241. https://doi.org/10.1016/B978-008044612-7/50064-0

[7] Kandi SG, Tehran MA, Hassani N, Jarrahi A. Color naming for the Persian language. *Color Res Appl*. 2015;40(4):352-360.

[8] Jameson KA, Webster MA. Color and culture: innovations and insights since basic color terms—their universality and evolution (1969). *Color Res Appl*. 2019;44(6):1034-1041.

[9] Hardin CL. Berlin and Kay theory. *Encyclopedia of Color Science and Technology*. Vol 1–4. Springer Science; 2013.

[10] Greenberg JH. *Universals of Language*. The M.I.T. Press; 1963.

[11] Siemund P. *Linguistic Universals and Language Variation*. De Gruyter Mouton; 2011.

[12] Croft W. Methods for FInding language universals in syntax. In: Scalise S, Magni E, Bisetto A, eds. *Universals of Language Today*. Springer; 2009:145-164.

[13] Ramat P. The (early) history of linguistic typology. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:1-10.

[14] Daniel M. Linguistic typology and the study of language. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:1-16.

[15] Comrie B. *Language Universals and Linguistic Typology*. The University of Chicago Press; 1989.

[16] Dryer MS. Large linguistic areas and language sampling. *Stud Lang*. 1989;13(2):257-292.

[17] Bakker D. Language sampling. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:1-26.

[18] Moravcsik EA. Explaining language universals. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:69-89.

[19] Epps P. Linguistic typology and language documentation. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:1-10.

[20] Dąbrowska E. What exactly is universal grammar, and has anyone seen it? *Front Psychol*. 2015;6:1-17.

[21] Dryer MS. Why statistical universals are better than absolute universals. *Papers from the 33rd Annual Meeting of the Chicago Linguistics Society*. Chicago Linguistics Society; 1998:1-23.

[22] Lehmann WP. *Syntactic Typology: Studies in the Phenomenology of Language*. University of Texas Press; 1978.

[23] Vennemann T. Analogy in generative grammar: the origin of word order. *Proceedings of the Eleventh International Congress of Linguists*. Il mulino; 1974:79-83.

[24] Vennemann T. Categorial grammar and the order of meaningful elements. In: Juliand A, ed. *Linguistic Studies Offered to Joseph Greenberg on the Occasion of his Sixtieth Birthday*. Anma Libri; 1976:615-634.
[25] Dryer MS. The Greenbergian word order correlations. *Language*. 1992;68(1):81-138.
[26] Hawkins JA. Implicational universals as predictors of word order change. *Language*. 1979;55(3):618-648.
[27] Hawkins JA. On implicational and distributional universals of word order. *J Ling*. 1980;16(2):193-235.
[28] Croft W. *Typology and Universals*. Cambridge University Press; 2003.
[29] Siewierska A. *Word Order Rules*. Croom Helm; 1988.
[30] Allan K. Hierarchies and the coince of left conjuncts (with particular attention to English). *J Ling*. 1987;23(1):51-77.
[31] Corbett G. Implicational hierarchies. In: Song JJ, ed. *The Oxford Handbook of Linguistic Typology*. Oxford University Press; 2010:1-13.
[32] Comrie B, Keenan EL. Noun phrase accessibility Revisted. *Language*. 1979;55(3):649-664.
[33] Smith-Stark TC. The plurality Split. In: La Galy MW, Fox RA, Bruck A, eds. *Papers from the Tenth Regional Meeting, Chicago Linguistic Society, April 19–21*. Chicago Linguistic Society; 1974:657-671.
[34] Corbett G. The agreement hierarchy. *J Ling*. 1979;15:203-224.
[35] Comrie B. Syntactic typology. In: Mairal R, Gil J, eds. *Linguistic Universals*. Cambridge University Press; 2006:130-154.
[36] Levinson SC. Yeli Dnye and the theory of basic color terms. *J Ling Anthropol*. 2000;10(1):3-55.
[37] Saunders B. Disinterring basic color terms: a study in the mystique of cognitivism. *Hist Hum Sci*. 1995;8(4):19-38.
[38] McCarthy AD, Wu W, Mueller A, Watson B, Yarowsky D. Modeling color terminology across thousands of languages. *EMNLP-IJCNLP 2019–2019* Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. 2020 2241–2250.
[39] Kay P, Cook RS. Word color survey. In: Luo MR, ed. *Encyclopedia of Color Science and Technology*. Springer; 2015:1-8.
[40] Kay P, Berlin B, Maffi L, Merrifield W. *Color Naming across Languages*. No. July 2013. Cambridge University Press; 2009.
[41] Regier T, Kay P, Cook RS. Focal colors are universal after all. *Proc Natl Acad Sci U S A*. 2005;102(23):8386-8391.
[42] Uusküla M. Basic colour terms in Finno-Ugric and Slavonic languages: Myths and facts. 2008 (August):206.
[43] Paggetti G, Menegaz G, Paramei GV. Color naming in Italian language. *Color Res Appl*. 2016;41(4):402-415.
[44] Paramei GV, Griber YA, Mylonas D. An online color naming experiment in Russian using Munsell color samples. *Color Res Appl*. 2018;43(3):358-374.
[45] Zaslavsky N, Garvin K, Kemp C, Tishby N, Regier T. The Evolution of Color Naming Reflects Pressure for Efficiency: Evidence from the Recent Past. 2021.
[46] Corbett G, Davies IRL. Linguistic and Behavioural measures for ranking basic colour terms. *Stud Lang*. 1995;19(2):301-357.
[47] Hays DG, Margolis E, Naroll R, Perkins DR. Color term salience. *Am Anthropol*. 1972;74:1107-1121.
[48] Johansson S, Hofland K. Frequency analysis of English vocabulary and grammar: based on the LOB corpus. *Volume I: Tag Frequencies and Word Frequencies*. Clarendon; 1989.
[49] McManus IC. Basic colour terms in literature. *Lang Speech*. 1983;26(3):247-252.
[50] McManus IC. Half-a-million basic colour words: Berlin and Kay and the usage of colour words in literature and science. *Perception*. 1997;26(3):367-370.
[51] Nivre J, De Marneffe MC, Ginter F, et al. Universal dependencies v1: a multilingual treebank collection. Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016. 2016 1659–1666.
[52] Machálek T. KonText: advanced and flexible corpus query Interface. *Proc LREC*. 2020;2020:7005-7010.
[53] Perkins RD. Statistical techniques for determining language sample size. *Stud Lang*. 1989;13(2):293-315.
[54] Cysouw M. Using the world atlas of language structures. *Lang Typology Univ*. 2009;61(3):1-6.
[55] Odden D. Languages and universals. *J Univers Lang*. 2003;4(1):33-74.
[56] Hammarstrom H. Counting languages in dialect continua using the criterion of mutual intelligibility. *J Quant Ling*. 2008;15(1):34-45.
[57] Mattiola S. *Two Language Samples for Maximizing Linguistic Variety*. AMS Acta; 2020.
[58] Miestamo M, Bakker D, Arppe A. Sampling for variety. *Ling Typol*. 2016;20(2):233-296.
[59] Majid A, Levinson SC. WEIRD Languages Have Misled us, Too. 2010.
[60] O'Keeffe A, McCarthy M. *The Routledge Handbook of Corpus Linguistics*. Routledge; 2013.
[61] Kamholz D, Pool J, Colowick SM. PanLex: Building a resource for panlingual lexical translation. *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014. 2014 3145–3150.
[62] Abdramanova S. Basic color terms in the Kazakh language. *SAGE Open*. 2017;7(2):215824401771482.
[63] Mylonas D, Griffin LD. Coherence of achromatic, primary and basic classes of colour categories. *Vis Res*. 2020;175:14-22.

AUTHOR BIOGRAPHIES

Antoni Brosa Rodríguez is a Martí i Franquès researcher at Universitat Rovira i Virgili. He works on universals in language, through the relationship between linguistic typology, fuzzy logic and computational linguistics.

M. Dolores Jiménez López is a professor at Universitat Rovira i Virgili and main researcher of the Research Group on Mathematical Linguistics. The application of formal models to natural language analysis is one of her main research interests.

How to cite this article: Brosa-Rodríguez A, Jiménez-López MD. Quantifying basic colors' salience from cross-linguistic corpora. *Color Res Appl*. 2024;49(1):34-50. doi:10.1002/col.22899

APPENDIX A

TABLE A1 Analyzed languages.

Language	Tokens	Indoeuropean	Language	Tokens	Indoeuropean
Afrikaans	49 K	Yes	Indonesian	220 K	No
Akkadian	2 K	No	Irish	3 M	Yes
Armenian	23 K	Yes	Italian	117 M	Yes
Basque	121 K	No	Japanese	184 K	No
Belarusian	8 K	Yes	Kazakh	10 K	No
Breton	10 K	Yes	Kurmanji	10 K	Yes
Bulgarian	73 M	Yes	Latin	2 M	Yes
Buryat	10 K	No	Latvian	93 M	Yes
Catalan	920 K	Yes	Lithuanian	96 M	Yes
Chinese	260 K	No	Maltese	296 K	No
Croatian	28 M	Yes	Norwegian	856 K	Yes
Czech	910 M	Yes	Old Church Slavonic	58 K	Yes
Danish	88 M	Yes	Persian	55 K	Yes
Dutch	97 M	Yes	Polish	100 M	Yes
English	4G	Yes	Portuguese	125 M	Yes
Ezyra	16 K	No	Romanian	74 M	Yes
Estonian	79 M	No	Russian	99 K	Yes
Faroese	10 K	Yes	Slovak	1G	Yes
Finnish	72 M	No	Slovenian	101 M	Yes
French	133 M	Yes	Spanish	114 M	Yes
Old French	171 K	Yes	Swedish	86 M	Yes
Galician	126 K	Yes	Thai	22 K	No
German	94 M	Yes	Turkish	362 K	No
Gothic	55 K	Yes	Uighur	40K	No
Greek	98 M	Yes	Ukranian	116 K	Yes
Ancient Greek	420 K	Yes	Upper Sorbian	11 K	Yes
Hebrew	250 K	No	Urdu	95 M	Yes
Hindi	5 M	Yes	Vietnamese	44 K	No
Hungarian	93 M	No			

TABLE A2 Hierarchy based on frequencies IPM.

B&K order	Clean (46)	Full (57)
Black	209.6	330.34
White	144.75	184.36
Red	109.58	136.24
Green	77.53	88.18
Blue	57.99	63.02
Yellow	38.44	50.67
Gray	19.64	17.59
Purple	19.12	15.43
Brown	12.66	22.49
Pink	12.19	13.28
Orange	8.19	10.57

TABLE A3 Comparison in German.

Colors	Our results	Hays et al.
White	1st	1st
Red	2nd	2nd
Black	3rd	3rd
Green	4th	4th
Blue	5th	5th
Yellow	6th	6th
Brown	6th	6th
Gray	6th	6th
Pink	9th	9th
Orange	10th	9th
Purple	11th	9th