

What's in a Colour? Studying and Contrasting Colours with COMPARA

Diana Santos¹, Rosário Silva², Susana Inácio²

Linguatca, COMPARA: ¹SINTEF ICT, Norway; ²FCCN, Portugal

Diana.Santos@sintef.no, mrosariomsilva@netcabo.pt, susana.inacio@fccn.pt

Abstract

In this paper we present contrastive colour studies done using COMPARA, the largest edited parallel corpus in the world (as far as we know). The studies were the result of semantic annotation of the corpus in this domain. We chose to start with colour because it is a relatively contained lexical category and the subject of many arguments in linguistics. We begin by explaining the criteria involved in the annotation process, not only for the colour categories but also for the colour groups created in order to do finer-grained analyses, presenting also some quantitative data regarding these categories and groups. We proceed to compare the two languages according to the diversity of available lexical items, morphological and syntactic properties, and then try to understand the translation of colour. We end by explaining how any user who wants to do serious studies using the corpus can collaborate in enhancing the corpus and making their semantic annotations widely available as well.

1. Motivation

Colour is one of the semantic domains more used in linguistic argumentation over the status of language vs. cognition, and language vs. world. See for example Berlin & Kay's (1969) influential work, and its support or criticism in works by Pinker (1994) or Sampson (2005), or Gärdenfors (2000). It is also a prime example of morphological creativity, diachronic change and cultural differences. See e.g. Cheminée et al. (2006) for the latter. This is why we started our semantic explorations of COMPARA by the colour domain. As far as we know, COMPARA (Frankenberg-Garcia and Santos, 2003)¹ is the largest edited parallel corpus in the world, and after the revision of its automatic syntactic annotation (for nouns and adjectives; verbs are under way) in Portuguese (Santos and Inácio, 2006; Inácio and Santos, in progress), and the beginning of the same process in the English side, we thought it was high time to use it for semantic studies as well.

2. The Annotation Process

Since there are currently no available automatic semantic analysers that we know of, we decided to use a lexically-driven approach followed by human revision. After having manually extracted all the words denoting colour from several lists of words automatically extracted from COMPARA, we marked automatically all colour words in COMPARA² on both sides (4,774 in Portuguese and 4,808 in English), amounting to 428 Portuguese different forms (279 different lemmas) and 483 different forms in English. This was encoded in the *sem* attribute for query purposes (we use CQP (Christ et al., 1999; Evert, 2005) for corpus encoding in the DISPARA system; see Santos (2002) for more details). So, to retrieve colour instances in COMPARA the query should be [*sem*="cor.*"] or [*sem*="colour.*"], for

Portuguese and English, respectively. Then, all cases were manually reviewed, and, in order to give an idea of all the potential meanings colour can present, we separated those instances where it was clear that colour denoted human race, by further specifying *cor:raça/colour:race* in both languages. We also clustered in a new category for Portuguese words that deal primarily with hair and skin, marked as *cor:humana*. For colour words related to wine we further employed the label *cor:vinho*.

For English, we also assigned to the words *red* and *white* the classification *colour:wine* when they referred to wine. The category chosen to classify human colours was however considerably broader than its Portuguese counterpart, in that it spanned over much more than just skin or hair colour. This is probably due to the fact that many "coloured" words in English are compounds such as *red-eyed*, *red-whiskered* and so on, which would be rendered as two different words in Portuguese.

The use of *verde* and *green* in the two languages to convey the meaning of unripe was also marked in the attribute *sem*, as *naomaduro* and *unripe* respectively.

Finally, we have separately marked (using the attribute *cor:original/colour:original*), words that originally referred to colour but whose main meaning has gone well beyond the mere colour reference. Examples are *blue* in *Blue Danube*, *yellow* in *Yellow pages*, and words like *greyhound*, *greenhouse*, *ovelha negra* (black sheep) or *sorriso amarelo* (grimace).

In Tables 1 and 2 we show the quantitative distribution of all these cases in COMPARA, for both languages (originals and translations), or just for originals. The underscore joins vague cases, i.e., cases where the annotators did not feel they had enough information to decide between. Vagueness is something that has been consistently marked in COMPARA also in other contexts, like PoS or morphological number, see Santos (1998;2006) for a defence of it. 0 means not colour at all, while *cor_naomaduro* indicates those cases where it is not clear whether the occurrence denotes something with a green colour or just unripe, or both.

The whole annotation process is thoroughly documented

¹ www.linguatca.pt/COMPARA/

² We have used the version 10.0.2 of COMPARA for the data presented in this paper, with 1.5 million words in each language.

in Silve et al. (in progress).

Attribute	Port	Eng
cor/colour	3,412	3,505
cor:raça/colour:race	676	641
cor:humana/colour:human	371	276
cor:vinho/colour:wine	16	16
cor:original/colour:original	161	260
cor_cor:raça/colour_colour:race	6	43
cor_cor:humana/colour_colour:human	2	1
cor_cor:original/colour_colour:original	88	8
cor_0/colour_0	35	51
cor_naomaduro/colour_unripe	5	6

Table 1: Colour categories in PT and EN (originals and translations).

By this initial comparison we can see a large difference in the use of words belonging to the `colour:original` category. However, this difference is somewhat softened if we take into consideration the number of vague instances (vagueness between colour and `colour:original`). Another line of Table 1 that stands out is `cor_0/colour_0` (vagueness between colour and not colour) where we can see much more instances in English (51) than in Portuguese (35). An example in English is "...*blue face*" that can indicate sadness or change in the colour of the skin. In Portuguese a good example is "...*o verde da natureza*" which can mean all the greenery existing in nature (not necessarily green) or just the range of the green colour nature presents.

By comparing the two languages in its entirety, we observe that in Portuguese there are slightly more colour words related to race than in English, and significantly more human words in Portuguese. These two differences can be largely explained by the fact that we did not consider *Negro* a colour word in English, and by the fact that apparently the skin/hair field in Portuguese is more developed, as we will discuss in section 5. But also words like *morena* are rendered by *dark*, which denote shades in English and we did not mark as colour.

Attribute	Port	Engl
cor/colour	1,642	1,837
cor:raça/colour:race	326	383
cor:humana/colour:human	129	173
cor:vinho/colour:wine	5	12
cor:original/colour:original	67	185
cor_cor:raça/colour_colour:race	2	9
cor_cor:humana/colour_colour:human	1	-
cor_cor:original/colour_colour:original	44	4
cor_0/colour_0	22	22
cor_naomaduro/colour_unripe	4	1

Table 2: Colour categories in EN and PT (source texts).

Table 2 shows the distribution of the colour categories, but only for the original texts. Now, instead of representing the language used to describe the same states of affairs (assuming that the two versions of the texts represent roughly the same stories), we can compare what (the particular set of) English original texts attend to as opposed to (the other particular set of) Portuguese original texts. Interestingly, we note that English-speaking authors use more words that were classified into the categories `colour:race` and `colour:human` than Portuguese-speaking authors. Also curious is the imbalance regarding the category `colour:original`. Apparently words in English that have a colour reference in their origin but have gone beyond the mere colour sense are much more frequent. (This may however be also due to different subjective criteria of the different annotators, and has to be investigated further.)

2.1. Specifying the Colour Category Further

We also created manually specific attributes to classify colour words, i.e., we grouped all straightforward colour words into seventeen colour groups for each language -- see below -- in order to do finer-grained comparisons between authors and languages. The colour groups are Black/Preto and White/Branco (representing the outermost points of the colour spectrum), Blue/Azul, Yellow/Amarelo, Red/Vermelho, Orange/Laranja, Green/Verde, Purple/Roxo, Brown/Castanho, Beige/Creme, Pink/Rosa, Grey/Cinzeno, Gold/Dourado and Silver/Prateado (representing the most significant and predominant primary and secondary colours), Multiple/Multipla (to classify compound nouns where one unique colour reference cannot be identified, for instance, *grey-blue* and *yellow-red*; not for cases where the main colour reference is explicit, like in *bluish-green* or *greenish-yellow*), Unspecified/Naoespecificada (to indicate colours that are implicitly present in the text but not explicitly named, and to mark the cases in which there is reference to many colours as in for example, *colourful* or *different-coloured*). Finally, we used the group Other/Outras to encompass all colours that do not fall easily into other categories (or about which there was no consensus in the annotation team)..

Examples of members of each group (encoded in the attributes `cor` or `colour` for Portuguese and English respectively) can be seen in Tables 3 and 4,. Note that this grouping is independent of part of speech.

Those tables show the size (i.e., the number of different words – or types – belonging to the group), and the extent (number of instances of colour words in that group in the whole corpus), for each group as well. They are organized in decreasing order of group size: Vermelho and Red are the simple colours with most different forms (48 and 35) in the two languages, while Laranja and Orange are have the least variety (6 and 4). The main differences in the rankings of the two languages, in fact, occur in the complex categories (Multiple, Other and Unspecified)

and are apparently due to non-semantic factors, such as spelling differences or different morphological features).

Group	Size	Example of words belonging to the group	Extent
Vermelho	48	encarnado, purpúra, purpureamente, ...	427
Azul	34	azuis-pálidas, azula, anilada, turquesa, ...	321
Branco	30	alvo, branco-sujo, esbranquiçava, ...	644
Verde	32	esverdeado, verde-azulados, ...	271
Amarelo	27	amareleciam, amarelo-esverdeada, ...	178
Cinzeno	27	plúmbeos, cinza, cinza-claro, cor-de-rato, ...	126
Castanho	24	acastanhada, marrom, castanho-rosada, ...	142
Outras	22	pardo, fulvos, bronze, âmbar, ocre, malva, ...	123
Preto	17	negro, enegrecido, pretume, negríssimo, ...	509
Dourado	16	ouro, dourava, doiradas, douradura, ...	153
Rosa	16	cor-de-rosa, rosado, róseos, rosa-shoking, ...	122
Naoespecificada	15	cor, coloridas, multicolor, coloriam, variegado, ...	353
Roxo	15	violeta, lilás, arroxou, violáceo, ...	70
Prateado	9	platinados, prateando, platina, ...	34
Múltipla	9	verde-negra, rosa-pérola, marmorizado, ...	10
Creme	8	beges, pérola, marfim-velho, ...	29
Laranja	6	cor de laranja, alaranjado, ...	33

Table 3: Colour groups in Portuguese.

3. Initial Exploratory Studies

COMPARA can be used, like its source of inspiration, the English Norwegian Parallel Corpus, ENPC (Johansson & Hofland, 1994), (i) to compare two languages, (ii) to compare original and translated text in the very same language, and (iii) to compare what happens when things are translated in either direction. Furthermore, and due to the presence of several different varieties of both Portuguese and English, it can also be used as a tool for (iv) comparing varieties (provided, in all cases, that we are aware of the size of the material and do not extrapolate too wildly).

Since COMPARA includes texts from a variety of authors, it can also be used for literary studies and for comparing

different styles. In Inácio et al. (2008) and Silva et al. (2008) the different behaviour among English-speaking authors and Portuguese-speaking ones as far as colour was concerned was contrasted and discussed.

We also tried to assess whether there was a consistent pattern of increase in colour use with time, but the results were inconclusive.

Group	Size	Example of words belonging to the group	Extent
Multiple	38	black-and-white, grey-green, marbled, ...	57
Red	35	scarlet, reddish, crimson, ruby, brick-red, ...	436
Green	34	greenish, olive, emerald, pale-green, ...	265
White	30	snow-white, whitened, white-collared, ...	633
Blue	29	pale-blue, sky-blue, aquamarine, blue-checked, ...	323
Grey	29	gray, greyed, pearl-grey, ashen, ...	171
Unspecified	28	colour, colored, colourful, multicoloured, ...	353
Brown	28	auburn, muddy-coloured, mustard-brown, ...	189
Pink	23	rose-colored, pinkish, dusty-pink, ...	126
Black	21	black, blackened, coal-black, ...	487
Yellow	21	yellowed, yellowish, ...	177
Other	14	bronze, amber, ochre, copper, sand-colored, ...	41
Gold	9	golden, ginger-gold, ...	145
Purple	9	violet, lilac, violet-coloured, ...	95
Beige	7	cream, ivory, pearl-coloured, ...	33
Silver	5	silvered, silvering, ...	40
Orange	4	orange-colored, carrot-colored, ...	33

Table 4: Colour groups in English.

In table 5, we compare the colour groups of both source languages. Interestingly, Pink, Grey, Brown and Orange are more frequent in English originals, while Verde (Green), Preto (Black), Branco (White) and Amarelo (Yellow) are more common in Portuguese.

The morphological productivity of hyphens in English allows several multiple colours to be created, such as *black-and-white*, *blue-green*, or *rose-red*, which leads to a higher number of Multiple cases. On the other hand the use of the word *cor* (colour in Portuguese) without any further specification is undoubtedly more frequent in

Portuguese than in English.

Another interesting observation is that the colour words belonging to the groups Dourado (Gold) and Prateado (Silver) are far more frequent in Portuguese than in English, given that the words *gold* and *silver* are often classified as *cor_0* (vague between colour and not colour), due to their double sense. In fact, *gold* in *gold letters* or in *gold taps* may as well indicate something that has the colour of gold or which is made of gold itself.

EN (original)		f	PT (original)		f
White	266	32.5	Branco	365	57.1
Black	246	30.1	Preto	263	39.2
Blue	167	20.4	Azul	152	23.8
Yellow	96	11.7	Amarelo	84	13.1
Red	240	29.3	Vermelho	186	29.1
Orange	22	2.7	Laranja	11	1.7
Green	129	15.8	Verde	139	21.7
Purple	40	4.9	Roxo	47	7.3
Brown	130	15.9	Castanho	32	5
Beige	24	2.9	Creme	10	1.6
Grey	112	13.7	Cinzentos	39	6.1
Pink	98	12	Rosa	29	4.5
Gold	65	7.9	Dourado	71	11.1
Silver	23	2.8	Prateado	21	3.3
Other	23	2.8	Outras	64	10
Multiple	47	5.7	Múltipla	5	0.8
Unspecified	143	17.5	Nãoespecificada	197	30.8

Table 5: Distribution of colour groups in English and Portuguese source texts. *f* stands for the frequency for each one hundred thousand words

4. Morphosyntactic Analysis

We also did some morphosyntactic studies, for example investigating the percentages with which the colour adjective position is postposed, preposed or predicated in relation to the noun it modifies; the use of relative clauses modifying colour; and coordination in Portuguese (Inácio et al., 2008) and then also comparing with English (Silva et al., 2008).

We established that in COMPARA the more common coordinated colour (group)s are Branco, Vermelho, Azul and Preto in Portuguese originals, while White, Pink, Red and Blue are most frequent in English originals.

Morphological Patterns	PT	EN
Hyphenated words	219	335
"X-colo(u)red" / "Cor de ..."	105	61

Table 6: Morphological patterns of colours.

We have also investigated morphological patterns such as

hyphenated compounds and the expression *cor de X* (corresponding to the English *X-coloured*).

Kind of word	Freq	Example
colour-[borrowed from something]	85	cherry-red
colour-[hue modifier]	66	deep-blue
colour-[applied to something]	66	green-shirted
colour-["colo(u)red"]	61	carrot-colored
colour-[colour]	46	pink-brown
colour-[hair-related]	44	red-haired
colour-[skin-related]	20	olive-skinned
colour-[eye-related]	8	gray-eyed

Table 7: kinds of hyphenated colour words in English

Table 6 presents the colour distribution in COMPARA of these morphological patterns, while Table 7 furnishes an overview of the kinds of hyphenated words in English. Table 8 presents the colour distribution in COMPARA by part of speech in the two languages. Nouns are conspicuously more frequent in Portuguese, due to a much more tenuous boundary than in English between the two categories adjectives and nouns (which have for example the same inflectional paradigms). The opposite can be seen for verbs in English. Surprisingly, there is a marked preference for proper nouns including colours in Portuguese as compared to English (but given that the English parsing has not yet been humanly revised this may be the explanation).

Colour Part-of-Speech	PT	EN
Adjectives	3,201	3,721
Nouns	1,326	932
Proper Nouns	106	40
Verbs	90	108
Adverbs	1	7

Table 8: Grammatical category of colours.

5. Contrastive Analysis

The main goal of the present paper is to investigate the way colour is translated, or not translated, into English and Portuguese.

Comparing the colours in both languages (original and translated texts), we obtain the data of Table 9. The column labelled *Differences* shows, for instance, the number of cases in which Branco was not translated by White, next to the cases in which White was not translated by Branco, 14% and 12%, respectively.

The groups showing most correspondence are Laranja, Azul, Rosa, Castanho and Amarelo. The groups showing most differences are Outras and Múltipla. Given that the group Other is much smaller in English, it is obvious that there cannot be a good correspondence of the two groups in the two languages. Likewise, the same applies to the Multiple group, which is much larger in English.

