CorTrad and Portuguese-English translation studies: investigating colours

Diana Santos, University of Oslo and Linguateca
Stella E. O. Tagnin, Department of Modern Languages, University of São Paulo
Elisa Duarte Teixeira, CoMET Project, University of São Paulo

Abstract

CorTrad is a bidirectional and multiversion English-Portuguese parallel corpus. By multiversion we mean that it consists of more than one version of the translated texts. It is also semantically tagged for colours and clothing. This paper will report on some exploratory studies on translations of colours using CorTrad. Colours are of special interest for translation because they refer to visual properties, and have a strong metaphorical import. Moreover, not all colour metaphors transfer across English and Portuguese, which make their study relevant also for translation teaching. Some of the questions that guided our investigation are: a) How are colours used in English and Portuguese?; b) How often do colours have a terminological, metaphorical, or figurative content?; c) How are colours translated?; d) Are there different translation equivalents for the “same” colour?; and e) How important is genre for colour translation?. Our findings so far reveal that: i) technical texts use descriptive colours very sparsely, although colours are pervasive in technical terminology; and ii) colours used in technical terminology vary across languages, increasing the chances of translation pitfalls.

1. Introduction

Colours are of special interest for translation because they refer to visual properties, and have a strong metaphorical import (Philip 2011). Moreover, not all colour metaphors transfer across English and Portuguese (Santos et al. 2008). In this paper we will address the following questions: a) How are colours used in English and Portuguese?; b) How often do colours have a terminological, metaphorical or figurative content?; c) How are colours translated?; d) Are there different translation equivalents for the “same” colour?; and e) How important is genre for colour translation? To this end, section 2 presents a short description of the CorTrad corpus and its three subcorpora. Section 3 explains briefly the system and rules used to annotate colour and clothing. A discussion of choices made in relation to problematic cases for the annotation is carried out in section 4. Section 5 focuses on the actual findings for colour. Our concluding remarks are in section 6.

2. The CorTrad corpus

CorTrad is a parallel, multiversion corpus for the Portuguese-English language pair. We call it multi-version because some of CorTrad’s texts include more than one version of the translated text, as explained in the next paragraph. CorTrad is a collaborative enterprise between two projects: CoMET, the Multilingual Corpus for Teaching and Translation, hosted by the Department of Modern Languages of the University of São Paulo, Brazil, and DISPARA, hosted by Linguateca, a distributed language resource centre for Portuguese. CorTrad has been introduced to the international community in Tagnin, Teixeira & Santos (2009) and to the Portuguese-speaking community in Teixeira, Santos & Tagnin (2012). Access to the corpus is free, and anyone can use it in their own research, as well as inspect our annotation choices and validate or criticize our conclusions.

CorTrad is currently made up of three subcorpora: **CorTrad literário**, a fiction/short story subcorpus with original texts in English and three translation stages into Portuguese (version 1: translation-trainee texts; version 2: revised translations; and version 3: published texts). At the time of writing, it consists of 28 Australian short stories; 20 Canadian short stories will be included shortly. **CorTrad técnico-científico** hosts a Brazilian cookbook translated from Portuguese into English by Brazilian translators (version 1), and then revised by a native American English speaker (version 2). And **CorTrad jornalístico** features a set of 1,072 scientific magazine articles translated from Portuguese into English in Brazil. Work on enlarging CorTrad content is under way.

CorTrad is built on DISPARA (Santos 2002) and is syntactically annotated by PALAVRAS (Bick 2000) for Portuguese, and POS-annotated by CLAWS (Rayson & Garside 1998) for English. From Linguateca's viewpoint, CorTrad is a follow-up to the AC/DC (Access and Availability of Corpora) project and, in fact, it is often described under the larger term "AC/DC cluster", since it shares most of its options and processing features with the Portuguese corpora of AC/DC (Santos 2011). Incidentally, this setup realizes Stig Johansson's model for parallel corpus compilation, since translated texts are part of a larger monolingual pool (Johansson & Hofland 1994).

All texts have been semantically tagged for colour and clothing in Portuguese, and so far only for colour in English, using the **Corte-e-Costura** ("made-to-measure" or "customized") system (Santos & Mota 2010). The tagged output has been manually revised. To the best of our knowledge, this is one of the first semantically tagged bilingual corpora available, which makes it relevant for a new generation of contrastive studies.

Also, given that the corpus was built, from the beginning, to encompass several versions of the "same" translation, specific functionalities have been devised not only to compare the two languages but also the multiple versions, as shown later in this paper. Thus, CorTrad is innovative in both content and setup.

3. The annotation process in a nutshell

CorTrad's semantic annotation process follows these steps:

1. Identifying a set of words (lemmas) associated with a semantic field;
2. Creating a set of rules that:
    1. correct cases of homonymy (as in example *de cor* below),
    2. provide annotation for some rare cases (which can only be considered colour in special contexts),
    3. remove annotations in other uncommon cases (that is, cases where most of the time the expression means colour, but not always),
3. Correcting and improving particularities of specific texts which are hard to generalize.

In step three, linguists review and rewrite the "rules" that only apply to very specific cases, usually bound to a specific text. The process of refining these rules (or commands) goes on until all cases are correctly annotated.

There are two levels of rules: general rules (which are shared by all AC/DC corpora – see again Santos (2011) –, and which should hold for most Portuguese texts), and text/corpus-specific rules based on a particular set of texts and cases. Note that, methodologically, we concur with Smith et al. (2008) in that we devise systems that help human annotation of a corpus instead of requiring labour-intensive editing. In case other texts are added or changes are made to the texts (a case in point is a new version of the parsers), it should be more straightforward to inspect the changes or differences and reformulate the necessary rules, than editing the corpus itself.

To give an idea of how the annotation process takes place (see also Freitas et al. 2012), we present below some examples of rules involving colour in the specific syntax of the Corte-e-Costura program (all rules are available from the Linguateca website, specifically from the rules page):

```
a:[lema="oração"] b:[lema="de"] c:[lema="cor"] >> c:[sema="0"]
[lema="visão"] a:[lema="cor-de-rosa"] >> a:[sema="cor:original"] [1]
[word="tinto"] [lema="ou"] a:[lema="branco"] >> a:[sema="cor:vinho"]
```

The first rule removes the colour annotation from the phrase *(saber) oração de cor* ([know] a prayer by heart), where *cor* does not mean colour (it is a homograph). [2] The next two examples provide more specific categories for colour words in specific contexts, namely *visão cor-de-rosa* (“rosy view”), used in a metaphorical way, and *tinto ou branco* (red or white) in which “white” stands for a type of wine and not the colour white.

4. Annotation choices

Saying that colour was annotated in the corpus may seem trivial to a layperson, but whenever one is interpreting and describing language, there is a myriad of options and choices that need to be made. Extensive guidelines are in place for Portuguese (for an almost exhaustive report concerning colour annotation, see Silva & Santos 2012, in Portuguese). Here we report on the general approach:

- Multi-word expressions are marked with an `mwe` tag, so that it is possible to retrieve them as units, <mwe>yellow pages</mwe>, <mwe>black beans</mwe>.
- Only colour, not intensity, is tagged, so that for *dark blue* only the word “blue” is tagged.
- Adverbs, nouns and verbs, if related to colour, are also tagged, such as *blue-eyedly*, *blacks*, and *colouring*, but verbs that only presuppose colour, such as *paint* or *dye*, are not tagged.
- Colours expressed by more than one word such as *black and white* and *blue-greenish* are tagged (and counted) as one colour, as well as descriptive colours such as *the colour of your eyes when you cry* (“colour” is tagged), or *pink colour* (“pink” is tagged).
- Mentions of a non-specified colour such as *multicolour*, *hair colour*, or the absence of colour, such as *colourless*, are also tagged, respectively, as belonging to the Unspecified and Absence groups.

More controversial, still, is the classification of “pure colour” vs. other (conventionalized) uses of colour, which are tagged as `cor:original` (based on the fact that, originally, the words meant colour, but that this is no longer their main use, see also note 1). We also annotate three special cases (`human`, for cases like *blond*, *redheaded*, *blush*, and *pale*; `race`; and `wine`), given that they are clearly distinct from the pure perceived colour and are, in a way, technical or specialized.

The more distinctions and fine-grained points one tries to address, the harder it is to decide in context, and in fact an interesting disagreement between the three authors (solved by a majority rule) was found in the technical cooking term *dourar* (literally “give / acquire a golden colour”), which was considered a pure colour term by two authors, while one of the authors identified it as a cooking term – so that it should be considered as derived (and therefore marked as `cor:original`). Although apparently a matter of little significance, *dourar* is the most frequent colour term in CorTrad técnico-científico (to date), and therefore a different annotation choice would produce rather different quantitative results (as we show below).

As to clothing, we have, so far, only distinguished real clothing from clothing used in a metaphorical or conventional sense, again labelled `roupa:original`, such as *black tie*. Contrary to colours, which are mostly adjectives, clothing annotation concerns almost only nouns or verbs. In any case, it was surprising to find that many idioms and other figurative language uses for clothes are operative both in Portuguese and in English. This topic deserves further investigation in the future.

It should be noted that the semantic annotation of clothing and colours is orthogonal, in the sense that a particular expression or occurrence can be tagged both as colour and as clothing, and, in addition, it can also be tagged `original` (conventional) in one or even both domains, as in the case of *white-collar crimes* or *capacetes azuis* (“blue helmets”, metonymically denoting UN forces), or *camisa amarela* (“yellow jersey”, the leading cyclist).

5. Colours in CorTrad

We annotated CorTrad for colours in both languages and clothing vocabulary in Portuguese, and revised the outcome. One should perhaps insist that very different figures would have been found if only lexically-based counts had been made: *orange*, *gold* and even *white*, most of the times, do not denote (pure) colours at all, and the same is true of clothing words such as *meia* (“socks” but a homonym of “half”), *capa* (“raincoat” but also meaning “book cover”) or *anel* (“ring”) and, in English, *ring*, which, in addition to referring to jewellery, is often used in contexts such as Saturn rings, bullfight rings, etc.

Looking at the results in each subcorpus, it is fair to state that colour tagging makes it easier to identify linguistic and cultural differences because we can explicitly look for cases of colour mismatch: for example, different connotations are known for *blue* in Germanic languages – related to depression and dark – and in Portuguese, related to good and bright. To discuss whether this could be an example of linguistic relativity is beyond the scope of this paper.

In the following sections, we present some examples of colour and clothing tagging by genre in CorTrad. A quantitative sum-up of colour, in the major categories, is shown in the top half of Table 1 below. (In this and other tables vague cases are counted once for every interpretation, that is, if the linguist could not decide between two or more interpretations, the occurrence is considered vague between them. Also, we always use the last version of the translation.) We remind readers that by “original” we mean cases where an expression, which originally denoted colour, took on other uses and meanings that became conventional, such as *green light*, *black hole* or *red herring*. [1] This also applies for English compounds such as *greenhouse* or *blackbird*.

	Cooking (source)	**Scientific news (source)**	**Short stories (translation)**
Size in words	134,093	768,721	74,869
Pure colour	584	341	316
“Original”	315	186	3
Race	0	56	12
Human	0	6	36
Wine	87	1	4
Total	578	589	373

Table 1. The distribution of colour and clothing in the Portuguese part of CorTrad.

The distribution for English is shown in Table 2.

	Cooking (translation)	**Scientific news (translation)**	**Short stories (source)**
Size in words	151,990	839,206	75,041
Pure colour	641	321	293
“Original”	825	255	27
Race	0	56	15
Human	0	3	12
Wine	104	1	4
Total	1,566	637	352

Table 2. The distribution of colour in the English side of CorTrad

5.1 Studying colours in the CorTrad corpus

The first step we took in the study of colours in CorTrad was to generate a list of the colour words in each corpus, in both languages. For that purpose, we typed the query `[sema="cor.*"]` in the search interface of the corpus and selected the option `Distribuição das formas` ("Distribution of forms") for each subcorpus. We did that for the original corpus and for the last version of the translated texts. Thus, although the data is presented by language below, in the Science News and in the Cookbook subcorpora, Portuguese is the language of the original texts (noted as `or.` from now on), whereas in the Short Story subcorpus, it is the language of the translated texts (marked as `tr.` from now on). The results are shown in Tables 3 and 4, where all lemmas with 5 or more occurrences are displayed.

Science news (or.) 83 lemmas / 125 forms		**Short stories (tr.) 65 lemmas / 109 forms**		**Cookbook (or.) 43 lemmas / 83 forms**	
Total occ.	589	Total occ.	373	Total occ.	991
cor	70	branco	55	dourar	341
branco	52	azul	41	branco	144
vermelho	43	vermelho	34	vermelho	122
verde	42	negro	24	verde	115
negro	38	cor	22	preto	39
buraco	29	preto	17	cor	38
infravermelho	29	amarelo	16	tinto	33
ultravioleta	28	verde	15	dourado	27
azul	25	pálido	14	amarelo	16
amarelinho	24	cinza	11	roxo	14
febre (amarela)	22	rosado	10	rosado	12
preto	21	marrom	9	esbranquiçado	6
amarelo	20	cor-de-rosa	6	alaranjar	6
transparente	14	cinzento	6	amarelado	6
colorido	8	grisalho	6	alaranjado	6
corante	8	colorir	5	colorir	6
avermelhar	6	amarelado	5	esbranquiçar	5
verdura	6	transparente	5	Other	4
avermelhado	5	dourado	5		
amarelado	5	Other	4		
colorir	5				
Other	4				

Table 3. Portuguese colour lemmas in CorTrad

Science News (tr.) 56 lemmas / 68 forms		**Short Stories (or.) 50 lemmas / 62 forms**		**Cookbook (tr.) 35 lemmas / 49 forms**	
Total occ.	637	Total occ.	352	Total occ.	1,566
black	99	white	61	brown	388
color	78	black	47	black	263
white	74	blue	44	white	259
green	67	red	41	red	219
red	62	grey	22	green	204

yellow	51	brown	18	color	56
blue	44	green	18	golden	41
yellowing	25	yellow	15	yellow	26
greenhouse	20	colour	13	purple	16
gray	15	pink	11	pink	12
golden	8	Other	4	browned	11
colored	7			blue	10
blackboard	7			greenish	7
reddish	6			orange	7
milky	6			yellowish	6
brown	6			beige	5
pink	6			colored	5
Other	4			Other	4

Table 4. English colour lemmas in CorTrad

Pure colour (obtained with the search expression `[sema="cor"]`) was also investigated, given that we expected that purely visual uses of colours would be less used in technical text, and the results are shown in Tables 5 and 6. It should be stressed that PALAVRAS assigns the verbal lemma to adjectival uses of past participles. Therefore, cases like *esbranquiçado* ("whitish") or *colorido* ("coloured") appear in the lemma list respectively as *esbranquiçar* and *colorir*.

Science news (or.) 59 lemmas / 91 forms		**Short stories (tr.) 49 lemmas / 88 forms**		**Cookbook (or.) 31 lemmas / 64 forms**	
Total occ.	341	Total occ.	313	Total occ.	578
cor	70	branco	40	dourar	341
vermelho	40	azul	39	verde	60
verde	40	vermelho	33	cor	38
branco	27	cor	22	branco	28
azul	22	negro	18	dourado	26
amarelo	19	preto	16	rosado	12
preto	15	amarelo	16	amarelo	7
transparente	13	verde	13	esbranquiçado	6
colorido	8	pálido	14	amarelado	6
avermelhar	6	cinza	11	colorir	6
avermelhado	5	marrom	9	esbranquiçar	5
amarelado	5	rosado	7	Other	4
colorir	5	cinzento	6		
Other	4	colorir	5		
		amarelado	5		
		cor-de-rosa	5		
		dourado	5		
		transparente	5		

		Other	4		

Table 5. Portuguese pure colour lemmas in CorTrad

Science News (tr.) 43 lemmas / 49 forms		**Short Stories (or.) 42 lemmas / 52 forms**		**Cookbook (tr.) 30 lemmas / 36 forms**	
Total occ.	321	Total occ.	293	Total occ.	637
color	73	white	44	brown	318
green	41	blue	42	color	56
red	37	red	38	green	55
yellow	25	black	35	white	52
blue	25	grey	17	golden	31
white	22	green	16	red	22
black	19	yellow	15	yellow	17
colored	7	colour	13	browned	11
reddish	6	brown	12	greenish	7
gray	6	pink	11	orange	7
milky	6	Other	4	purple	6
brown	5			yellowish	6
Other	4			black	6
				pink	5
				beige	5
				colored	5
				Other	4

Table 6. English pure colour lemmas in CorTrad

One can see at once from the tables above that the three subcorpora show quite a different distribution. The density of colours is much higher in fiction, and the percentage of pure colours compared to colours used in other more technical or conventionalized senses is, again, much higher in fiction, and quite low in the other subcorpora. Finally, the variety of the colour expressions is much higher in the scientific news, which cover a broad spectrum of subjects, in comparison to the cookbook corpus which focusses on a single theme: cooking. Also worth noticing is the fact that each domain shows a different set of colour terms which are the most frequent ones. The next step was to look at each subcorpus individually.

5.1.1 The Cooking domain

Analysing the topmost occurring colours in the original and translated texts of the Cookbook corpus, one notes that the prima facie colour equivalent for *dour.**, *golden*, had only 41 occurrences in English (as compared to 341 occurrences of the verb *dourar*, plus 27 occurrences of the adjective *dourado* in Portuguese). This rendered *golden* the 5th most frequent lemma, as opposed to the lemma *dourar*, which was the most frequent lemma in Portuguese. We then wondered: if *dour.** is the most frequent in the Portuguese list of CorTrad técnico-científico (i.e. the cookbook sub-corpus), why does *brown* (and not *golden*) show up as the most frequent form in the English list of colours?

We did a basic search for the lemma *dour.**, which returned 368 occurrences. A brief look at the concordance lines (a few examples are reproduced below) revealed some interesting translation solutions that may not occur to a novice or non-specialized translator:

Ex.	Original	Revised Translation
1	Leve ao forno por mais ou menos 1 hora e 15 minutos, mexendo de vez em quando, até que os pedaços estejam *bem* **dourados** e macios.	Bake for approximately 1 hour and 15 minutes, turning occasionally, until vegetables are *deep* **golden brown** and fork-tender.
2	Aqueça um fio de azeite numa frigideira grande e **doure** *ligeiramente* a cebola.	Heat a drizzle of olive oil in a large frying pan and *lightly* **brown** the onion.
3	Espalhe sobre o espinafre e leve ao forno por 30 minutos, até formar uma crosta **dourada**.	Sprinkle mixture evenly over the spinach and bake for 30 minutes, until a **golden** crust forms.
4	Misture o pistache, o orégano, o sal, o azeite e a pimenta numa assadeira média, espalhe e leve ao forno por 10 a 15 minutos, até que esteja **dourado** e crocante.	Combine pistachio nuts, oregano, salt, olive oil, and black pepper in a medium-sized baking pan, spread evenly and bake for 10 to 15 minutes, until they are **golden** and crunchy.
5	Abaixe a temperatura para 200º C (médio-alto) e asse por uns 25 minutos, até que os pães estejam crescidos e **dourados**.	Lower oven temperature to 200C (400F / moderately hot / Gas 6) and bake for 25 minutes or so, until bread loaves are risen and **golden brown**.
6	Para as almôndegas, **doure** *ligeiramente* os pinoli numa frigideira seca e reserve.	For the meatballs, *lightly* **toast** pine nuts in a dry frying pan and set aside.
7	Aqueça um fio de azeite numa frigideira grande, junte metade das almôndegas, deixe **dourar** primeiro de um lado e depois dos outros.	Heat a little olive oil in a large frying pan, add half the meatballs and fry, turning once halfway through, until thoroughly **golden brown**.
8	Para o tomate, aqueça um fio de azeite numa panela média e **doure** *ligeiramente* a cebola.	For the tomato sauce, heat a drizzle of olive oil in a medium pot and *lightly* **brown** the onion.

Table 7. Examples of concordance lines returned for the search of `dour.*` in CorTrad's Cookbook corpus. [3]

Also, we realized that the word *golden* occurred 218 times, out of which 177 times in the phrase *golden brown*. (Note that this is marked as colour under *brown* – cf. the remarks above on *dark blue*.) The adjective *dourado* is translated as *golden brown* in examples 1 and 5, and only as *golden* in examples 3 and 4. As a verb, usually in the imperative (*doure*) or infinitive (*dourar*), it is translated with the verb *brown* (examples 2 and 8), *toast* (example 6), and changed into an adjectival phrase (*golden brown*) in example 7. Whether this may be considered a lexical gap in English (there is no verb *golden*), the fact that *brown* and *dourar* are equivalents in cooking texts has an impact on translation practice and training. Also worthy of notice is that the intensifier *bem* is rendered as *deep* in English (example 1), while *ligeiramente* becomes *lightly* (examples 2, 6 and 8).

We may in fact suggest that the use of *dourar* in Brazilian Portuguese (interestingly, in Portugal, another verb, *alourar*, is more frequently used in that context) is the reason for the possibly excessive frequency of the verb *to brown* and the expression *golden brown* in the English translation, where a native speaker might have employed less visual cooking terms like *fry* or *toast* in many cases. This remains to be verified in larger cooking corpora, though, and is not supported by the intuition of the majority of the authors.

It was also interesting to observe (see Table 8) that cooking ingredients tend to have names including colour, but in some cases the colours are different in each language.

Ex	Portuguese (or.)	English (tr.)
9	vinho tinto [=*tinted*]	red wine
10	ovos vermelhos [=*red*]	brown eggs
11	uva passa clara [=*light coloured*]	golden raisins
12	repolho roxo [=*purple*]	red cabbage

Table 8. Examples of colours translated by a different colour or term

There are many cases where ingredients that have no colour word in Portuguese are translated by ingredients with colour words in English, as Table 9 shows. The opposite may well be true, although we have no native English materials for cooking in CorTrad to compare with. We did find *frutas vermelhas* (“red fruits”) rendered as *berries*, though.

Ex	Portuguese (or.)	English (tr.)
13	açúcar mascavo	brown [=*marrom*] sugar
14	pão de forma	white [=*branco*] bread
15	pimenta-do-reino	black [=*preta*] pepper
16	ervilha (comum)	green [=*verde*] peas

Table 9. Examples of non-colour Portuguese terms translated by an English term including colour

Note that it is not possible to provide a literal translation for the Portuguese terms in most cases. *Mascavo*, for instance, is a word that only appears in this cooking term, having no other uses in current Portuguese. For the sake of clarity, we add the word that, in Portuguese, would correspond to the colour in English, like *brown=marrom*. There are other cases where colours are used differently, such as for describing parts of ingredients, as in the following example in Table 10.

Ex	Portuguese (or.)	English (tr.)
17	Escorra, descarte as partes mais **duras** [=*hard*] e fibrosas e corte em tiras de 1 cm.	Drain, discard any **dark green** or fuzzy portions and slice into 1 cm (3/8 in) strips.

Table 10. Example of non colour formulations that are translated by colour ones

5.1.2 The Fiction genre

In the investigation of fiction, we looked for colour mismatches. In what follows, we present and discuss some of our findings.

For one thing, colour is not always translated as a colour, as Table 11 illustrates.

Ex	English (or.)	Portuguese (tr.)
18	... just declared itself out to the **blue**	... surgira de repente, do nada (appeared suddenly, from nowhere)
19	It was to be **black** tie.	O traje é a rigor.
20	She never refused to go to Melbourne, but it was her hoodoo city, a **black** jinx.	Nunca se negava a ir a Melbourne, mas era uma cidade de azar, mau agouro.

21	The dog knew they were coming, and barked **blue** murder.	O cachorro sabia que eles estavam vindo e latiu desesperadamente (desperately).
22	(...) would quarrel with her till the **white** hours	(...) de discutir com ela até o amanhecer (dawn)
23	(...) knowing about **brown** rice	(...) entendia de arroz integral
24	had no more idea of its value than any average bush **blackfellow**.	mas não tinha idéia do seu valor, não mais que qualquer outro aborígine.

Table 11. Examples of colours that are not translated as colours

Examples (18)–(22) are actually idiomatic expressions, sometimes translated by equivalent expressions in Portuguese (19, 20), sometimes made explicit (18, 21, 22). Examples (23) and (24) are considered terms in their domains and are translated by their equivalent terminology. None of them features a colour in the target language, though.

Occasionally colours are dropped altogether, maybe because translators deemed them unnecessary (examples (25), (26), (27) and (29)), or different associations are created (examples (30)).

Ex	English (or.)	Portuguese (tr.)
25	She rushed to the back of the house and hauled the drowsy **black** [=*preto*] pup out of the kennel.	Ela correu para os fundos da casa e puxou o sonolento filhote para fora da casinha.
26	She began to prowl between the desks, waving the **white** [=*branca*] letter like a flag.	Então começou a rondar as carteiras, balançando a carta como se fosse uma bandeira.
27	When the torrent of **white** water subsided	Quando a corrente de águas espumantes [=*foamy*] cessou,
28	The **brown** [=*marrom*] smell of his cigar	O aroma característico do charuto
29	**white** and blue enamel bowls	tigelas azuis [=*blue*] esmaltadas
30	She had the **tinted** view of the Irish	Ela tinha a visão **cor-de-rosa** [=*pink*] dos irlandeses

Table 12. Examples of colours that are omitted in translation, or given a different colour

5.1.3 The Scientific News genre

Another interesting finding is the large number of colour terms in scientific terminology found in CorTrad jornalístico. This ranges from names of diseases and description of symptoms in medical or agricultural texts to stars and other astronomical phenomena whose names include colours, through descriptions of racial conditions in demography or sociology, and the use of colours in computer-aided geography and cartography. A further source of colour terminology involves botanical species. This is incidentally also the case in the cookbook material, where a high percentage of ingredients have a colour specifier (*pimenta branca*, *carnes vermelhas*, *feijão preto*, etc., that is: literally white pepper, red meat, black beans).

Given that in both genres (technical and scientific news) the presence of a colour word in technical terms is often related to their specification (from a common genus), these may be translated by a different colour or term, as the following examples in Table 13 show.

	Portuguese (or.)	English (tr.)
31	Ali vivem outras oito espécies ou subespécies de aves endêmicas: o jacamim (Psophia viridis interjecta), um periquito (Pyrrhura perlata anerythra), a mãe-da-taoca (Phlegopsis confinis), o **papa-formiga** (Pyriglena leuconota interposita), uma **araponga** (Procnias alba wallacei), o **chupa-dente** (Conopophaga aurita pallida) ...	It is there that another eight species or subspecies of endemic birds live: the dark-winged trumpeter (Psophia viridis interjecta), Neumann's pearly conure (Pyrrhura perlata anerythra), a bare-eye (Phlegopsis confinis), the **white-backed fire-eye** (Pyriglena leuconota interposita), the **white bellbird** (Procnias alba wallacei), the **chestnut-belted gnateater** (Conopophaga aurita pallida) ...
32	A equipe do Inpa simulou vazamentos de óleo em aquários e analisou o impacto dessa agressão ambiental em exemplares de pirarucu e de dois tipos de respiradores aquáticos com adaptações diferenciadas à falta de oxigênio (hipoxia), o boari (Mesonauta insignis) e o **tambaqui** (Colossoma macropomum).	The team from Inpa simulated oil spills in aquariums and analyzed the impact of this aggression against the environment on specimens of arapaima and of two kinds of aquatic breathers with differentiated adaptations to the lack of oxygen (hypoxia), the boari (Mesonauta insignis) and the tambaqui or **black pacu** (Colossoma macropomum).
33	Ali não se encontram mais, por exemplo, a anta (Tapirus terrestris) e o **porco-do-mato ou queixada** (Tayassu pecari), ainda que a floresta de Una seja grande o suficiente para abrigá-los.	For example, there are no more tapirs (Tapirus terrestris) and **white lipped peccaries** (Tayassu pecari), even though the Una forest is big enough to be home to them.
34	A **arara-canindé** (Ara ararauna), embora considerada extinta em São Paulo, é comum nos buritizais e nos cerrados da América do Sul, do Paraguai ao Panamá.	The **blue and gold macaw** (Ara ararauna), although being regarded as extinct in São Paulo, is common in the land where the mauritia palm grows and in the savannas of South America, from Paraguay to Panama.
35	Hoje, o **veado-campeiro** ocorre basicamente na parte central do Pantanal, não existe no sul e está pouco presente no norte.	Nowadays, the **red deer** exist essentially in the center of the Pantanal; there are none in the south, and very few in the north.
36	Entre as atuais espécies encontradas no Pará e que constam na lista oficial de animais brasileiros ameaçados de extinção estão a **ararajuba**, o **guará**, a arara-azul-grande, o sagüi-branco e o cuxiú-de-nariz-branco, entre outros.	Among the species currently found in Pará and which are on the official list of Brazilian animals threatened with extinction are the **golden parakeet**, the **scarlet ibis**, the hyacinth macaw, the **black-tailed marmoset**, and the white-nosed saki, amongst others.
37	"Está tudo no limite máximo, tanto a captura industrial do **camarão-rosa** (espécies do génreo Penaeus), da piramutaba(Brachyplatystoma vaillantti) e do **pargo** (Pagrus pagrus) quanto a artesanal da pescada-amarela e do guarijuba (Tachysurus luniscutis)", ...	"It is all at the upper limit, both the industrial catch of the **pink shrimp** (a species of the Penaeus genus), the laulao catfish (Brachyplatystoma vaillantti) and of the **red porgy** (Pagrus pagrus), and the rudimentary fishing of the acoupa weakfish and of the 'guarijuba' (Tachysurus luniscutis)",...
38	E há frutos que nem precisam de animais para dispersar suas sementes: usam o vento para isso, pois são dotados de asas ou plumas, como o araribá (Centrolobium tomentosum), o **jequitibá** (Cariniana legalis) e o cedro (Cedrella fissilis).	There are also fruit that do not require animals for the dispersion of their seeds. They use the wind for this as they are endowed with wings or feathers, such as the araribá (Centrolobium tomentosum),

		the **red bramble palm tree** (Cariniana legalis) and the cedar (Cedrella fissilis).

Table 13. Examples of scientific terms which use colour for naming of species in English, but not in Portuguese

We believe that, when terms are created through translation (in the specific scientific or technical area concerned), the same colour or a “standard translation” for the colour in the original is used; for example, in the case of *camarão rosa* and *pink shrimp* in example (37), where *rosa* and *pink* are prima facie equivalents. On the other hand, when they originate independently in different cultures or languages, there is room for a different conceptualization (and thus a different colour or no colour at all).

In the next section, we present a comparison of some aspects of colour vocabulary across the three subcorpora.

5.1.4 Comparing the three subcorpora

From the list of the most frequent colours in all subcorpora, we extracted their collocates, to see if they varied across corpora. Table 14 shows: a) the number of occurrences of the word *white* in each CorTrad subcorpus (e.g. 59 times in the Science News corpus); b) the number of lemmas represented by these occurrences (e.g. 36 lemmas in the Science News corpus); and c) the collocates immediately to the right of white in English texts (translated English in the case of Science News and Cookbook subcorpora) with two or more occurrences in each subcorpus.

Science News (tr.) 59 occ. / 36 lemmas		**Short stories (or.) 44 occ. / 26 lemmas**		**Cookbook (tr.) 41 occ. / 37 lemmas**	
dwarf	7	man	8	wine	51
cube	3	hand	7	chocolate	15
house	3	feather	5	rice	14
spot	3	hair	2	part	14
blood	3	Other	1	pith	9
shrimp	2			pepper	9
stripe	2			sandwich	8
hair	2			button	5
man	2			bean	5
cell	2			hominy	4
milk	2			smoke	4
paper	2			custard	4
light	2			bread	4
book	2			truffle	2
Other	1			sesame	2
				sea	2
				radish	2
				carrot	2
				variety	2
				grape	2
				meat	2

				Other	1

Table 14. Number of occurrences, lemmas and collocates of *white* in CorTrad

Table 15 shows the same information for the Portuguese texts. However, the query extracted the most frequent collocates immediately to the left of the lemma *branco* (the prima facie equivalent of *white* in Portuguese), as adjectives are usually positioned to the right of nouns in this language.

Science News (or.) 45 occ. / 31 lemmas		**Short stories (tr.) 42 occ. / 23 lemmas**		**Cookbook (or.) 122 occ. / 22 lemmas**	
glóbulo	4	mão	9	vinho	45
cabelo	4	homem	8	parte	14
cubo	3	pena	6	chocolate	12
mancha	3	cabelo	5	arroz	9
pelagem	2	látex	2	pele	7
luz	2	galão	2	pão de fôrma	7
população	2	blusa	2	milho	5
célula	2	Other	1	pimenta-do-reino	3
Other	1			carne	3
				peixe	3
				batata-doce	2
				louça	2
				Other	1

Table 15. Number of occurrences, lemmas and collocates of lemma *branco* in CorTrad

Another possibility of the CorTrad annotation and the DISPARA system is to look for colour groups instead of individual lemmas, since there are many distinct lexical forms per colour group. [4] So, a similar comparison could be obtained by queries such as `[grupo="Verde"]` or `[grupo="Green"]` and ask for lemma distribution.

5.2 Interaction between colours and clothing

By annotating the same corpora with two different semantic domains (for Portuguese only so far), we are able to study the ways in which they interact.

First, we observed that there is a significant co-occurrence of clothing and colours in fiction, while this is not the case in the other genres. When describing clothes in fiction one tends to also describe their colour. In Table 16, where “colour sentences” are those which have one or more cases of colour, and the same for clothing, we list the number of sentences with both colour and clothing annotation. A chi-square test shows that their co-occurrence is significantly higher than chance would have produced.

	Colour sentences	**Clothing sentences**	**Colour & clothing**	**Neither**
Fiction	290	228	60	4,683
Cooking	849	14	2	7,181
Scientific news	465	52	3	29,465

Table 16. Interaction between the domains

As might be expected, this correlation occurs in fiction only; it does not exist in cooking nor in scientific news.

Lexico-syntactically, colour and clothing behave quite differently, as colour is mainly adjectival, while clothing is mainly nominal. Although not shown in the present paper, both have a sizeable number of associated verbal expressions and seem to be prone to conventionalization and fixedness.

Studying both semantic domains allowed us to identify that colour is a salient property of clothing and is often employed in fiction, not only because their co-occurrence is higher than chance, but also because they enter into several conventionalized and highly cultural expressions. It remains to be studied whether other semantic domains (such as animals, plants or buildings) have a similarly strong correlation with colour, in fiction or in other genres.

6. Concluding remarks

We presented CorTrad as a good resource for studying authentic language in use, as well as authentic translation. The tools offered to explore the corpus enable users to see how different people, in different contexts, cope with the many challenges of translating colour (and clothing) vocabulary in different genres and for different target audiences.

We cannot state for the moment how specific the examples presented here might be or how easy it would be to generalize about the conclusions arrived at. For that reason, we plan to compare the data presented here with other semantically annotated monolingual and parallel corpora to verify whether our findings can be substantiated by comparable studies.

Although the corpora investigated may be too small for generalizations, we believe that the annotated and revised CorTrad corpus can be quite useful in teaching the two languages and teaching translation between them. Additionally, CorTrad can also be used as a source of interesting new hypotheses for further investigation, using other corpora. We hope that, beyond our results, our methodology may be of interest to others, and that comparable studies in other domains (and corpora) will benefit from our work.

Our future plans involve carrying out a more in-depth observation of the clothing annotation and its relation to colours. We also plan to annotate CorTrad for the food domain, that is, to mark all words and expressions referring to food in the corpus, to see how much attention is given to the domain in texts which are not about food or cooking. In this way we might be able to identify metaphorical and nontechnical uses in the semantic domain of food. In addition, we envisage to have CorTrad annotated also for the fear domain (Maia & Santos, this volume) in the near future, so that we can further investigate the relationship of colour and clothing with fear.

Acknowledgments

Thanks to Eckhard Bick and Paul Rayson, for the use of PALAVRAS and CLAWS, respectively.

Thanks to Sandra Aluísio and Arnaldo Candido Júnior at NILC for hosting the CorTrad corpus at NILC and for their corresponding technical support.

The work by the first author has been partially funded by the Linguateca project, which has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), UMIC, FCCN and FCT.

Notes

[1] This tag was an unfortunate terminological choice, made a while ago in the COMPARA project, and which, due to legacy issues, not only in the abundant documentation but also in the many rule files and colour annotated corpora in Linguateca, still has to await a global fix. Readers please have in mind that "original colour" is meant in the sense of "no longer only pure colour". Alternative descriptions might be "innovative", "conventional", "not mainly colour".

[2] Not always, but in the quite frequent metaphorical use, as in *Although the health authorities have not yet given the green light for the use of the equipment for psychiatric diagnosis, [...]*

[3] The first version of the translation was omitted because no differences were observed in the colour terms as compared to the revised translation.

[4] A colour group contains all colour terms that were considered to represent the same physical colour, using the common-sensical views of colour and the most frequent term to represent it. For example, the "brown" group includes, in addition to *brown*, also *brownish*, *auburn*, *chestnut*, etc. The groups are available for inspection and are discussed in Silva & Santos (2012).

Sources

COMET: http://comet.fflch.usp.br

CorTrad: http://comet.fflch.usp.br/cortrad

Linguateca: http://www.linguateca.pt/

DISPARA: http://www.linguateca.pt/dispara/

Rules for the Corte-e-costura program: http://www.linguateca.pt/ACDC/

References

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Freitas, Cláudia, Diana Santos & Rosário Silva. Forthcoming 2012. "Corpos e cores: colorindo a descrição da língua portuguesa". *Atas do ELC2011*, ed. by Deise Dutra & Heliana Mello. Campinas, SP: Mercado de Letras.

Johansson, Stig & Knut Hofland. 1994. "Towards an English-Norwegian parallel corpus". *Creating and Using English Language Corpora, Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*, ed. by U. Fries, G. Tottie & P. Schneider, 25–37. Amsterdam & Atlanta: Rodopi.

Maia, Belinda & Diana Santos. 2012. "'Who's afraid of ... what?' - in English and Portuguese.". *Aspects of corpus linguistics: compilation, annotation, analysis*. (Studies in Variation, Contacts and Change in English 12). Helsinki: Research Unit for Variation, Contacts, and Change in English. http://www.helsinki.fi/varieng/journal/volumes/12/maia_santos/.

Philip, Gill. 2011. *Colouring Meaning: Collocation and Connotation in Figurative Language*. John Benjamins.

Rayson, Paul & Roger Garside. 1998. "The CLAWS Web Tagger". *ICAME Journal* 22: 121–123.

Santos, Diana. 2002. “DISPARA, a system for distributing parallel corpora on the Web”. *Advances in Natural Language Processing (PorTAL 2002)*, ed. by Nuno Mamede & Elisabete Ranchhod. Berlin & Heidelberg: Springer-Verlag. Lecture Notes in Artificial Intelligence 2389. 209–218.

Santos, Diana. 2011. “Linguateca’s infrastructure for Portuguese and how it allows the detailed study of language varieties”. *Language Variation Infrastructure*, ed. by J.B. Johannessen. OSLa: Oslo Studies in Language 3.2. 113–128

Santos, Diana & Cristina Mota. 2010. “Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora”. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias. European Language Resources Association. 1437–1444.

Santos, Diana, Maria do Rosário Silva & Susana Inácio. 2008. “What’s in a colour? Studying and contrasting colours with COMPARA”. *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, 26 May – 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Silva, Rosário & Diana Santos. 2012. “Arco-íris: notas sobre a anotação do campo semântico da cor em português”. Current version: 12 June 2012. http://www.linguateca.pt/acesso/ArcoIris.pdf

Smith, Nicholas, Paul Rayson & Sebastian Hoffmann. 2008. “Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists’ Manual Annotations”. *Literary and Linguistic Computing* 23(2): 163–180.

Soares da Silva, Augusto. 2011. “Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese”. *Advances in Cognitive Sociolinguistics*, ed. by Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman. Berlin & New York: Mouton. 41–83.

Tagnin, Stella E.O., Elisa Duarte Teixeira & Diana Santos. 2009. “CorTrad: a multiversion translation corpus for the Portuguese-English pair”. *Arena Romanistica* 4: 314–323.

Teixeira, Elisa D., Diana Santos & Stella E. O. Tagnin. 2012. “CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês”. *Caminhos da Linguística de Corpus*, ed. by Tania Shepherd, Tony Berber Sardinha & Marcia Veirano Pinto. Campinas: Mercado de Letras. 151–176.

Studies in Variation, Contacts and Change in English 12: Aspects of corpus linguistics: compilation, annotation, analysis
Article © 2012 Diana Santos, Stella E. O. Tagnin & Elisa Duarte Teixeira; series © 2007– VARIENG
Last updated 2012-12-05 by Joe McVeigh
Links updated 2017-01-10 by Linda Ravindrarajan