Colouring COMPARA
Contrastive and monolingual colour studies in English and Portuguese

Rosário Silva, Susana Inácio, Diana Santos

In this paper we present the Portuguese and English colour studies we made using COMPARA (Frankenberg-Garcia & Santos, 2003), the largest edited parallel corpus in the world (as far as we know), as well as some findings concerning contrastive analysis.

Being an everpresent element in our world and in our lives, colour was our first choice in exploring COMPARA semantically. Moreover, this apparently simple subject has already given rise to a lot of linguistic argumentation over the status of language vs. cognition and language vs. world.

We start by discussing the annotation process regardig colour, namely, what is colour, what semantic categories were defined and what colour groups were created. Briefly, we marked as colour all straightforward words conveying colour, regardless of word class (*blue, reddened, blackly,* etc.), words that convey colour but are not in themselves a colour (*blond, brunette, blushed,* etc.), words that have colours in them but have gone beyond the mere colour reference (*greyhound, Whitehall*, etc.), and the word *colour* itself and its derivations. (The absence of colour was not marked.) We then defined five semantic categories (`colour`, `colour:race`, `colour:human`, `colour:wine` and `colour:original`) and classified our "colourful" words accordingly, having in mind the context in which they appeared. The words that fell into the category (pure) `colour` were later grouped into seventeen groups, to allow us finer-grained comparisons between authors and languages (Blue, Red, Yellow, Green, Orange, Brown, Beige, Black, White, Grey, Pink, Purple, Gold, Silver, Other, Multiple and Unspecified). (Silva, Inácio and Santos, 2008)

Having done this, we were able to discover which were the favourite colours of English-speaking authors, who contributes the most and the least to these colour preferences, what colour categories dominate each author's writing, etc., etc. Since preferences by the Portuguese authors in COMPARA were also investigated, we were able to contrast the use of colour in the two languages.

In addition, and taking advantage of the fact that the Portuguese part of COMPARA is syntactically analysed (automatically by the PALAVRAS parser (Bick, 2000) and then manually revised and documented, see Santos & Inácio, 2006), we studied colour-related syntactical patterns in Portuguese, and also described the kinds of lexemes associated to colour in the texts.

Finally, we discuss a few cases where colour is not translated or is changed, ending with some comments on translation practice in the two directions (English to Portuguese translation and Portuguese to English translation). Our study may present some literary surprises, in the sense that our initial expectations of more coloured authors were not confirmed.

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

Frankenberg-Garcia, Ana & Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus", in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translation Education*, Manchester: St. Jerome Publishing, 2003, pp. 71-87.

Santos, Diana & Susana Inácio. "Annotating COMPARA, a grammar-aware parallel corpus". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1216-1221.

Silva, Rosário, Susana Inácio & Diana Santos. "Documentação da anotação relativa à cor no COMPARA". Continually updated. First version: 27 November 2007. Current version: 8 February 2008. http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf