Contrastive Linguistics and Computation: mapping linguistic differences between Brazilian Portuguese and European Portuguese for rule-based computational model in Terminological Database

Currently, technology companies have been looking for new models to solve linguistic problems in technological demands, bringing together Linguistics and Computing Science. This work aims to present how a contrastive study between two Portuguese variants can contribute to the adaptation of content (localization) to improve the computational solution of a Terminological Database. To this end, we will present a linguistic-computational experiment that seeks to map the linguistic differences between Brazilian Portuguese (hereinafter BP) and European Portuguese (hereinafter EP) to implement them in a company's Terminological Database.

Localization, a field of translation, expanded in the early 1980s with the need to adapt the content or service of technology companies to increase their sales (Schäler, 2008). Thus, localization can be defined as the adaptation of linguistic and cultural material between different locale (Bendi, 2021). One way of analyzing languages in localization task is based on Contrastive Linguistics. This theory aims to compare or contrast two or more languages to establish the similarities and differences between language systems (Ke, 2018; Krzeszowski, 1990). By comparing a language pair, it is possible to examine the languages and subsequently apply the findings in translation, localization, and computing, such as in rule-based computational models.

For the linguistic-computational experiment, we conducted a contrastive study of BP and EP to generate computational rules and apply them to a dataset. First, we mapped and described the linguistic differences (cf. Castilho, 2009; Kato, 2006), creating 17 computational rules, which represent the linguistic knowledge and are predetermined by humans to define how the machine will interpret and apply the information. After this step, we generated a dataset of the company's terminology in table format, containing two spreadsheets with the results of translations from BP to EP and vice-versa, each consisting of 9109 instances. The next step was to implement 4 of the 17 rules in the BP to EP translation of the terminology’s dataset.

In an initial analysis, with the application of 4 rules in 1925 instances, we noticed that the rules (i) numerals and (ii) pronoun placement showed no changes between BP and EP, since the dataset was composed of terms, containing neither numerals nor words with pronoun placement. The application of the rule for the use of silent consonants in EP showed 1116 variations, and the rule of contrastive case in noun complement showed 815 changes. In the case of silent consonants, we found a high frequency of variation, because in EP the consonants 'p', 'c' and 'n' are used before another consonant and in BP this rule is not applied, as in "análise em direto" (BP) and "análise em directo" (EP), or in "gestão do setor público" (BP) and "gestão do sector público" (EP). The rule of contrastive case in noun complement showed a high variation, because in BP it is made a semantic generalization of the name and in EP it is common to particularize the name, for example, "interface gráfica de utilizador " (BP) and "interface gráfico do utilizador" (EP), or "recomendação de colaborador" (BP) and "recomendação do colaborador" (EP).

With the mapping and computational implementation of the linguistic differences between EP and BP, it was possible to observe that the rules executed on the dataset indicate that the contrastive analysis helped in rule-based computational solutions for building computational models for terminological database since the model was able to identify the differences between the variants with the rules. In the future, it is planned to apply all the rules and work with other lexical differences between the variants by adjusting the model on the dataset.

References

Bendi, M. (2021). The Reception of Localized Content: A User-Centered Study of Localized Software in the Algerian Market. In: Desjardins, R.; Larsonneur, C.; Lacour, P.. (2021). When Translation Goes Digital: Case Studies and Critical Reflections. Switzerland: Palgrave Macmillan.

Castilho, A. (2009). A hora e a vez do português brasileiro. Available in: <http://www.museulinguaportuguesa.org.br/colunas.php>

Kato, M. (2006). Comparando o Português da América com o Português de Portugal e com outras línguas. Língua Portuguesa, Museu da Língua Portuguesa.

Ke, P. (2018). Contrastive Linguistics. China: Peking University Press and Springer.

Krzeszowski, T. (1990). Contrasting Languages: The Scope of Contrastive Linguistics. Berlin, New York: Mouton de Gruyter.

Schäler, R. (2008). Linguistic resources and localization. In: Rodrigo, E. (2008). Topics in Language Resources for Translation and Localisation. Amsterdam / Philadelphia: John Benjamins Publishing Company.