Searching a Mixed Corpus in the Light of the New Portuguese Orthographic Norm

Gracinda Carvalho\textsuperscript{1,2,3}, Isabel Falé\textsuperscript{1}, David Martins de Matos\textsuperscript{2,4}, and Vitor Rocio\textsuperscript{1,3}

\textsuperscript{1} Universidade Aberta, Rua da Escola Politécnica, 147 1269-001 Lisboa, Portugal \{gracindac,imsfale,vjr\}@uab.pt
\textsuperscript{2} L2F/INESC-ID Lisboa, Rua Alves Redol 9 1000-029 Lisboa, Portugal david.matos@inesc-id.pt
\textsuperscript{3} CITI - FCT/UNL
\textsuperscript{4} Instituto Superior Técnico/UTL

Abstract. A mixed corpus of Portuguese is one in which texts of different origins produce different spelling variants for the same word. A new norm, which will bring together the written texts produced both in Portugal and Brazil, giving then a more uniform orthography, has been effective since 2009, but what happens in the perspective of search, to corpora created before the norm came into practice, or within the transition period? Is the information they contain outdated and worthless? Do they need to be converted to the new norm? In the present work we analyse these questions.

1 Introduction

Searching a corpus based on a user’s information request requires the words used on the query to be matched against those of the corpus. In the present work we analyse search effectiveness in the case of a corpus containing different variants of Portuguese, which we call a mixed corpus. The two variants correspond to the variant used in Portugal (pt\textsubscript{PT}) and the variant used in Brazil (pt\textsubscript{BR}). If a query is made in one variant, all answers occurring in other variants are not retrieved.

For a number of decades, countries with Portuguese as official language, have been involved in negotiations with the purpose of uniformising the orthography of both variants. An agreement was reached in 1990, which became known as “Acordo Ortográfico da Língua Portuguesa 1990”, or Portuguese Language Orthographic Agreement 1990 (PLOA). The former orthographic norm was in use until 2009, when the adoption of the new norm was effective in countries covering the vast majority of the Portuguese native speakers. The new norm will bring together the written texts produced both in Portugal and Brazil, giving then a more uniform orthography, but what happens to the corpus that were created before the norm came into practice, or within the transition period? Are they

\textsuperscript{1} As well as African Portuguese speaking countries, Macau and East Timor.
worthless? Do they need to be converted to the new norm? In the present work we analyse these questions.

In the present text, we describe our motivation and existing related work in Sect. 2. Sect. 3 is dedicated to the orthography describing the PLOA 1990 and its effects. In Sect. 4 we describe our solutions for the treatment of multiple spellings in QA@CLEF corpus, and in Sect. 5 we analyse the impact of the PLOA and our solutions in this corpus. To finalise, in Sect. 6 we present our conclusions and future work.

2 Motivation and Related Work

For an application that has the main goal of facilitating the access to information, it is important to consider as many sources of information as possible. Therefore, for a given language, a restriction to a specific variant would represent that a lot of useful information would be left out. That was the idea behind the creation of the corpus for Portuguese of the Cross Language Evaluation Forum (CLEF), an international campaign to evaluate several information retrieval related tasks. This corpus was created in 2004, consisting of the newspaper articles from Portuguese newspaper “Público” and the Brazilian “ Folha de São Paulo”. This collection became known as “CHAVE” and was organized by Linguateca.²

The edition of the Portuguese Wikipedia was added to the collection in the 2007 edition, with a frozen version of the HTML edition of the Portuguese Wikipedia from November 2006. The usefulness of the mixed corpus is referred in [1], that attributes the improvement of the performance by the addition of the pt_BR variant corpus, while in [2] the good performance of the lexical tools designed for the pt_PT variant in the scope of the mixed corpus is emphasised.

Studies have been made on mixed corpora of written Portuguese, particularly to determine the degree of convergence between them. That is for instance the idea behind the creation of the CONDIV (CONVergence DIVergence) corpus, though this work focused on semantic aspects [3].

The current orthographic norm has also been subject of a study in which the freely available lexical resources were combined and used for variant detection and conversion [4].

However we have no knowledge of a prior study dedicated to determining the best conditions in which to make documents belonging to both variants, before and after the 1990 norm, relevant and searchable. That is the purpose of the current work, together with the description of the particular techniques we find most appropriate to each particular case.

3 Portuguese Orthography: PLOA 1990 Rules

Officially sanctioned negotiations have taken place with the objective of creating a uniform intercontinental spelling of the Portuguese language, which culminated

² http://www.linguateca.pt/chave/