DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Internal Research Reports

ISSN

Number 60 | April 2019

Revisiting ETDs in Languages Other Than Portuguese @ PUC-Rio

Ana Maria Beltran Pavani

PUC RIO

Internal Research Reports

Number 60 | April 2019

Revisiting ETDs in Languages Other Than Portuguese @ PUC-Rio

Ana Maria Beltran Pavani

CREDITS

Publisher:
MAXWELL / LAMBDA/CCPA/VRAc
Sistema Maxwell / Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos
http://www.maxwell.vrac.puc-rio.br/

Organizers:
Alexandre Street de Aguiar
Delberis Araújo Lima

Cover:
Ana Cristina Costa Ribeiro

This extended abstract corresponds to a presentation ETD2018 - 21th International Symposium on Electronic Theses and Dissertations hosted by the National Library of Taiwan in Taipei, Taiwan, in September 2018. It is available in open access at
https://etd2018.ncl.edu.tw/images/phocadownload/73_Ana_Pavani_Extended_Abstract_ETD_2018.pdf
as a link from the symposium program at
https://etd2018.ncl.edu.tw/en/important-info/program.

REVISITING ETDs IN LANGUAGES OTHER THAN PORTUGUESE @ PUC-Rio

Ana Pavani, Member IEEE
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil
apavani@puc-rio.br

Abstract: This work addresses the publishing of ETDs in languages other than Portuguese both at PUC-Rio and in Brazil, focusing English. Initially, data compiled and organized on published ETDs are presented. Then, accesses to PUC-Rio ETDs are examined to identify the differences due to the languages. Data were collected from the systems that host ETDs; the surprising result is that ETDs in English have less accesses than the ones in Portuguese.

***Keywords*:** published ETDs, ETDs in Portuguese, ETDs in English, accessed ETDs.

INTRODUCTION

This paper addresses the languages used for ETDs of graduate programs of PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro. Languages of ETDs are a concern, among many, related to finding lengthy works written in many languages (Cayabyab, 2015). They also appear for languages with different scripts making it difficult to retrieve and/or to check plagiarism (Abraham, 2015). Creators (authors) and publishers (universities) are concerned because finding ETDs can turn into accesses and, eventually, citations. This work presents the evolution of the numbers of ETDs in languages other than Portuguese and the accesses they have had from different countries. It also shows the scenario of Brazilian ETD programs that are members of BDTD – Biblioteca Digital de Teses e Dissertações (http://bdtd.ibict.br/) which is the National Consortium established by IBICT – Instituto Brasileiro de Informação em Ciencia e Tecnologia in 2001. It is a continuation of a work that has been going on for some years and whose initial results were presented by the author (2017). It has a broader scope by including other Brazilian institutions and offering data that were gathered for all PUC-Rio ETDs in each language and not only a set of works as in 2017. The results come from data that are generated by systems that host ETDs in Brazil and at PUC-Rio. System logs and results from searches were processed to generate tables and graphics. All data for 2018 were collected up to August 03.

OBJECTIVES AND DATA GATHERING

The objectives of this work are to examine ETDs in languages other than Portuguese both at PUC-Rio and at BDTD. Accesses are only analyzed for PUC-Rio because there are no such data available for BDTD. It is not limited to presenting an overview of universities trying to make their ETDs easier to be internationally read but also focuses on comparing PUC-Rio to other Brazilian insitutions. In terms of accesses, it aims at examining how effective ETDs in English have been in broadening the audience and increasing the number of readers.

This section is divided in subsections to present specific topics.

PORTUGUESE IN THE WORLD

According to Ethnologue (https://www.ethnologue.com/statistics/size), Portuguese is the 7th most spoken language in the world; it is the 3rd Western language in the ranking. There are native speakers in 15 countries and they outnumber 220 million. The four countries with the largest populations are Brazil, Angola, Mozambique and Portugal; Brazil has about 87% of the native speakers. Considering Open Repositories, 3,782 were listed on OpenDOAR (http://v2.sherpa.ac.uk/view/repository_visualisations/1.html). Among them, 175 (4.6%) have contents in Portuguese; there are 99 Open Repositories in Brazil and 57 in Portugal.

ETDs PRESENTED AND PUBLISHED @ PUC-Rio

The ETD program of PUC-Rio began in May 2000 when the first ETD was published. In August 2002, ETDs became mandatory to all graduate programs (M and D levels). On August 03, 2018, there were 10,056 ETDs on the IR – Sistema Maxwell (https://www.maxwell.vrac.puc-rio.br/). In 2008, ETDs in languages other than Portuguese started being accepted with no special request to publish. The study of ETDs not in Portuguese has a time frame from 2008 on. Figure 1 shows the time series of the percentages of ETDs in Portuguese (pt) and in English (en) at PUC-Rio.

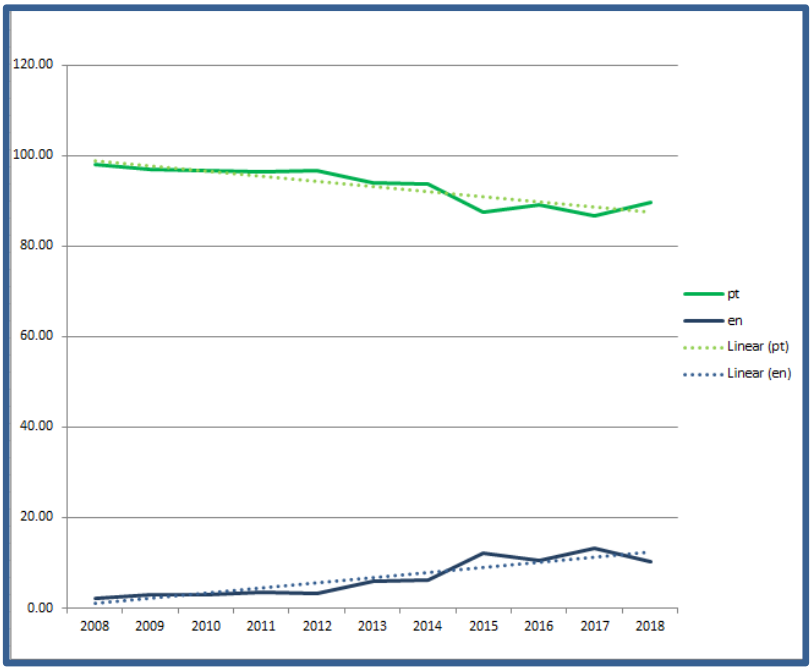

Figure 1. Time series of percentages of ETDs in Portuguese and in English at PUC-Rio from 2008 on.

Table 1 shows the percentages of ETDs in the two languages in two different time frames; the percentages in each time frme do not add to 100% because there are ETDs in French and in Spanish.

Table 1. Percentages of ETDs in English and in Portuguese in different time frames.

Time Frame	% pt	% en
All ETDs from 1966 on (10,056)	96.171	3.729
ETDs from 2008 on (6,424)	93.404	6.456

ETDs IN BRAZIL

Currently, there are 110 ETD programs that contribute to BDTD and the number of metadata records is almost 529K – one third at doctoral level. Some characteristics of the metadata records on the union catalog may be mentioned:

- USP – Universidade de São Paulo is the largest collection on BDTD with over 76K ETDs;
- There are 4 institutions whose programs have over 30K and less than 50K ETDs;
- PUC-Rio is in the third group – over 10K and less than 30K;
- There are 14 institutions whose programs have less than 100 ETDs.

Concerning the languages of ETDs, BDTD has 52 cooperating institutions with ETDs in English. There are ETDs in Spanish, French and Italian too, but they are not the focus of this work. Some characteristics of the numbers of ETDs in English on the catalog may be mentioned:

- Less than half of the member institutions have ETDs in English (52 in 110);
- Four institutions in 52 have between 5 and 10% of their ETDs in English;
- PUC-Rio has 3.729% of its ETDs in English;
- 35 institutions in 52 have less than 1% of their ETDs in English.

Figure 2 shows two histograms. The one on the left is of numbers of institutions per numbers of ETDs. The one on the right is of numbers of institutions per percentages of ETDs in English. Data used to create the histograms were gathered at (http://bdtd.ibict.br/vufind/Search/Results).

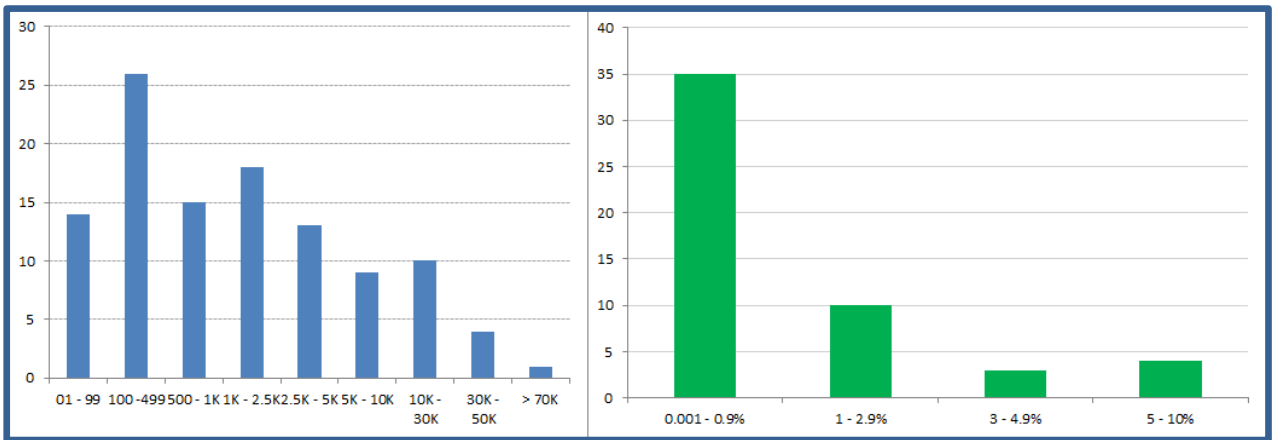

Figure 2. Numbers of institutions per sizes of ETDs collections and per percentages of ETDs in English on BDTD.

ACCESSES TO ETDs @ PUC-Rio

All contents on Sistema Maxwell are described and have sequential identification numbers. Since the system serves the institution, the types of contents are very diverse – there are ETDs, senior projects, monographs, articles, journals, books, hypermedia learning objects, simulator objects, research reports, technical reports, etc. ETDs account for almost 40% of all contents.

The system runs on CentOS, Apache, PHP and IBM DB2. In order to gather data on accesses, there are two automatic procedures that run every hour and they are fed by data recorded on the log of the Apache server. The first in for general accesses to the system and is not in the scope of this work. The second is for accesses to contents and is fully described in a previous work (Pavani, 2017). After its publication in 2017, a new set of programs was developed to yield

information on accesses to ETDs by language, i.e., by subcollections in different languages. Though access data have been collected since 2004 and could have been used, the decision was to analyze accesses in 2016, 2017 and 2018. The reason for the decision was the low number of ETDs in English before 2016. Data were arranged in the tables of a dataset (Pavani, 2018).

ANALYSIS OF RESULTS

Examination of the tables in the dataset (Pavani, 2018) indicates that:

- The patterns of accesses are very similar in the tables of columns 1 and 2 (ETDs in all languages and in Portuguese). This is easy to understand since the numbers of ETDs in Portuguese are much higher than the ones in English – the ratios between the numbers of ETDs in the two languages are approximately 44 (2016), 28 (2017) and 26 (2018). The ratio is going down – this is visible in figure 1.
- The average numbers of accesses to ETDs in Portuguese are much higher than to the ones in English.
  - The average numbers of accesses in 2016 are 131.74 (pt) and 21.91 (en).
  - The average numbers of accesses in 2017 are 146.62 (pt) and 18.04 (en).
  - The average numbers of accesses in 2018 are 87.89 (pt) and 10.88 (en).
- The average numbers of accesses from Brazil to ETDs in Portuguese are much higher than to the ones in English.
  - The average numbers of accesses in 2016 are 102.98 (pt) and 6.24 (en).
  - The average numbers of accesses in 2017 are 122.16 (pt) and 4.59 (en).
  - The average numbers of accesses in 2018 are 74.56 (pt) and 2.83 (en).
- Brazil is the country with most accesses regardless of the language of ETDs, as expected.
- The 10 countries with the highest numbers of accesses to ETDs in Portuguese include Mozambique , Portugal and Angola. They are not among the 10 with the most accesses to ETDs in English.
- Accesses from the United States, France and Germany, three countries in the top 10 in the two cases, are much lower for ETDs in English than in Portuguese.

The results are consistent in the three years that were examined. They indicate that works in English have significantly lower numbers of accesses than the ones in Portuguese. The reason for this seems to be the very low numbers from Mozambique, Portugal and Angola. Current results do not yield an interpretation of the fact that the United States, France and Germany rank among the countries with highest accesses for ETDs in Portuguese too.

CONCLUSIONS

The results on accesses are not very encouraging for publishing in English. There are two questions to be aswered in further studies: (1) Will higher percentages of works in English change the average accesses per work? (2) Are average accesses different for different subjects? To answer the first question more time is necessary. The second question will be aswered in the near future because programs are under development to yield these results.

The author would like to compare results with other ETD collections in Brazil – other institutions will be contacted to try a joint work.

REFERENCES

Abraham, Laila T. 2015. “Need for Language Technology in Developing ETD Packages”. ETD2015 Symposium in November 2015 in New Delhi, India. http://hdl.handle.net/10760/28325

Cayabyab, Terry A. C. 2015. “A Review of Emerging ETD Initiatives, Challenges and Future Developments”. *International Journal of Information and Education Technology* 5, No 10: 772-777. http://www.ijiet.org/papers/609-D031.pdf

Pavani, Ana. 2017. “ETDs in Languages Other Than Portuguese at PUC-Rio”. Paper presented and published on the website of ETD2017 Symposium in August 2017 in Washington DC. http://www.ocs.usetda.org/index.php/NDLTD/ETD2017/paper/viewFile/95/52

Pavani, Ana. 2018. “Accesses to ETDs in Portuguese and in English on Sistema Maxwell in 2016, 2017 and 2018 (up to August 03)”. Dataset. https://doi.org/10.17771/PUCRio.ResearchData.34785