DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Internal Research Reports

ISSN

Number 56 | April 2018

ETDs in Languages Other Than Portuguese at PUC-Rio

Ana Maria Beltran Pavani

PUC RIO

Internal Research Reports

Number 56 | April 2018

ETDs in Languages Other Than Portuguese at 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

ETDs in Languages Other Than Portuguese at PUC-Rio

Ana M B Pavani, *Member IEEE*
apavani@puc-rio.br
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brazil

Abstract: English is the international language to read on the Internet. It is also the language of the most prestigious journals for scholarly publications. Brazil is a large country with over 200 million inhabitants who speak Portuguese. ETDs in Brazil are mostly published in Portuguese thouch the articles that are their results most probably are in English. PUC-Rio has had an ETD program since 2000 and in 2002 they became mandatory. ETDs have been published in Portuguese, except for a few exceptions. In 2008, ETDs in languages other than Portuguese were formally allowed not to be exceptions. This work is the first to analyze and understand the consequences of allowing ETDs in foreign languages. It addresses the evolution of the numbers of publications in different languages and also the profiles of the accesses to the works, considering the countries of origin (language related) and numbers. Since in some cases the number of ETDs eligible for the analysis was very low, next year it will be repeated so that results can be compared.

Keywords: accessed ETDs; ETDs in English; ETDs in Portguese; published ETDs;

01. INTRODUCTION

This introduction has two objectives. The first is to examine the Portuguese language in the world – the numbers of native speakers and the countries where they live. The second is to introduce the context of the ETD program of Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

a. Portuguese in the World

Portuguese is among the 10 most spoken languages of the world. Various sources yield not the same information about the most spoken languages; some use data gathered in different years while others have to mix years due to the lack of information for some countries. Three sources were used though others were examined. The sources used were:

(1) Wikipedia – List of languages by number of native speakers (2007/2010) (https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers);
(2) Accredited Language Services – The 10 most common languages (2016) (https://www.accreditedlanguage.com/2016/09/09/the-10-most-common-languages/) ; and
(3) Ethnologue – Languages of the World (https://www.ethnologue.com/statistics/size) and Summary by Language Size (2017) (https://www.ethnologue.com/statistics/size) .

The information provided by the sources shows differences, probably due to years of collection of data. This difference does not impact the hierarchy among Western languages. Since Ethnologue [01] has the newest data and is also also a reference to other sites, the decision was to use its data. The 10 most spoken languages (in decreasing order of native speakers) are:

Chinese / Spanish / English / Arabic / Hindi / Bengali / Portuguese / Russian / Japanese / Lahnda

Regardless of the source used, Portuguese is always the third most spoken Western language, though in some it is the sixth or the seventh if Eastern languages are considered too.

The number of countries which have Portuguese as the official languages or one among others or as a used language exceeds 10. Except for Portugal, all are former Portuguese colonies. Since Portuguese sailors and colonists were all over the world, the countries are spreaad all over the world too. Table 01 shows the countries or autonomous regions that have Portuguese as one of the official languages. The contents of the columns are: (1) names of the countries; (2) continents where countries are located; (3) populations of the countries [02]; (04) UNDP Human Development Indices [03]; and (4) index I for each country [04].

Index I was created as an indicator of the potential to access ETDs since population and HDI do not individually represent such potential.

Country	Continent	HDI	Population	I
Angola	Africa	0.533	20,271,332	10,804,619.96
Brazil	South America	0.754	206,823,665	155,945,043.41
Cabo Verde	Africa	0.648	553,432	358,623.94
Equatorial Guinea	Africa	0.592	759,451	449,594.99
East Timor	Asia	0.605	1,261,072	762.948.56
Guinea-Bissau	Africa	0.424	1,757,159	745,035.41
Macau	Asia	0.894(*)	597,425	534,097.95
Mozambique	Africa	0.418	25,939,150	10,583,404.70
Portugal	Europe	0.843	10,833,816	9,132,906.89
Sao Tome e Principe	Africa	0.574	197,451	113,336.87
Total			**268,994,953**	**189,429,612.68**
Average		**0.629**	**26,899,495.3**	**18,942,961,27**

Table 1 – Countries, continents, HDIs, populations and indices I for Portuguese speaking countries.

(*) 2014 data collected from Wikipedia since the UN does not list Macau.

It is interesting to remark that Portugal that has approximately half of the population of Angola has index I only 10% lower. It is also important to observe the spread in the values of the HDIs – from 0.418 in Mozambique and 0.894 in Macau and 0.843 in Portugal. Macau is a Special Administrative Region of the People's Republic of China very close to Hong Kong (less than 60 miles).

HDI indicates life expectancy at birth, expected years of schooling, mean years of schooling and Gross National Income (GNI) per capita (UNDP – Human Development Index and its components at http://hdr.undp.org/sites/default/files/hdr_2016_statistical_annex.pdf). Schooling and GNI indicate the eagerness to access ETDs and the capability of doing it. In the case of capability it is possible to see two different aspects – the first is infrastructure to access the Internet which depends on the economic conditions of the country and the second is ability to understand the language of the works. In the case of this work, it is the ability to read and understand English which impacts accesses to ETDs in English as shown in section 03.

b. ETDs @ PUC-Rio

Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) started its ETD program in the year 2000. In August 2002, ETDs became mandatory to all graduate programs (M and D levels). Some

graduate programs digitized part or all traditional theses and dissertations, or are digitizing as time goes by. Currently, the ETD collection has over 9,000 ETDs – 78.3% are in open access. Many restricted ETDs come from the retrospective digitization when authors were not found to request authorization for their works to become available in open access.

Graduate programs are grouped in three areas: Humanities & Theology (8 active programs), Science & Technology (11 active programs) and Social Sciences (10 active programs). Their characteristics are very different since some of them are quite old (over 50 years) and others a very new (less than 5 years). The oldest and most traditional programs are in Science & Technology, where graduate studies began in 1963.

Up to until almost 10 years ago, theses and dissertations had to be written and published in Portuguese (pt-BR). For this reason, the number of works in this language is very high.

In 2008, other languages started to be allowed. Before 2008, only six ETDs were in foreign languages – five in English and one in Spanish, but they were exceptions. There is no information concerning theses and dissertations in printed format that were not digitized and added to the repository – the Maxwell System (https://www.maxwell.vrac.puc-rio.br/). When other languages started being allowed, the profiles of their adoption in the three areas were different and in all surprising low.

This work presents and comments the numbers of ETDs in foreign languages; data are shown. This is done in section 02. Since the time frame is from 2008 on, all are in digital format since ETDs have been mandatory since 2002.

Examining accesses to ETDs has been a focus at PUC-Rio for a long time. The gathering of data concerning accesses to ETDs in foreign languages (English as explained later on) and the comparison with analogous data for ETDs in Portuguese are presented in section 03.

Section 04 comments the results.

02. ETDs IN DIFFERENT LANGUAGES @ PUC-Rio

The current number (July 31, 2017) of published ETDs is 9,140. Among these, 4,838 were published in or after 2008, when ETDs in foreign languages started being accepted not as exceptions. Table 2 shows the percentages of ETDs in the four languages that are currently used.

Time Frame	% pt	% en	% es	% fr
All ETDs from 1966 on (9,140)	96.838	3.063	0.077	0.022
ETDs from 2008 on (4,866)	94.184	5.651	0.123	0.041

Table 2 – Percentages of ETDs in four different languages – considering the complete collection and the works published in the repository in 2008 and after.

The first line of table 2 contains all ETDs from 1966 on, this means it includes 2008 and after. Examination of raw data shows that before 2008, only five ETDs were published in English and one in Spanish. This represents 0.14% of all works published from 1966 to 2007.

The second line of table 2 is quite surprising. PUC-Rio has a very large number of graduate students from different countries in Latin America – all Spanish speaking. The author expected to find many works in Spanish and this is not the case.

It is quite obvious that the percentage of works in foreign languages has grown, but not as much as desired though. At the same time, the time series in figure 1 shows that the percentage is growing. The trend lines are clearly ascending for English and descending for Portuguese; but the rates are low. The lowest point of the Portuguese time series is 86.07% in 2017, at least considering the ones published until July 31. If this trend is maintained, the expectation is to have more ETDs in English as time goes by. Percentages for Spanish and French were not plotted because they are insignficant.

Some numbers may change due to delays in submissions by the authors and/or in the publishing and reviewing process. Since percentages are considered, variations are expected to be small. The trendlines for the numbers of ETDs in Portuguese and English are, respectively, decreasing and increasing. The numbers also indicate that the last 4 years seem to consolidate the trends.

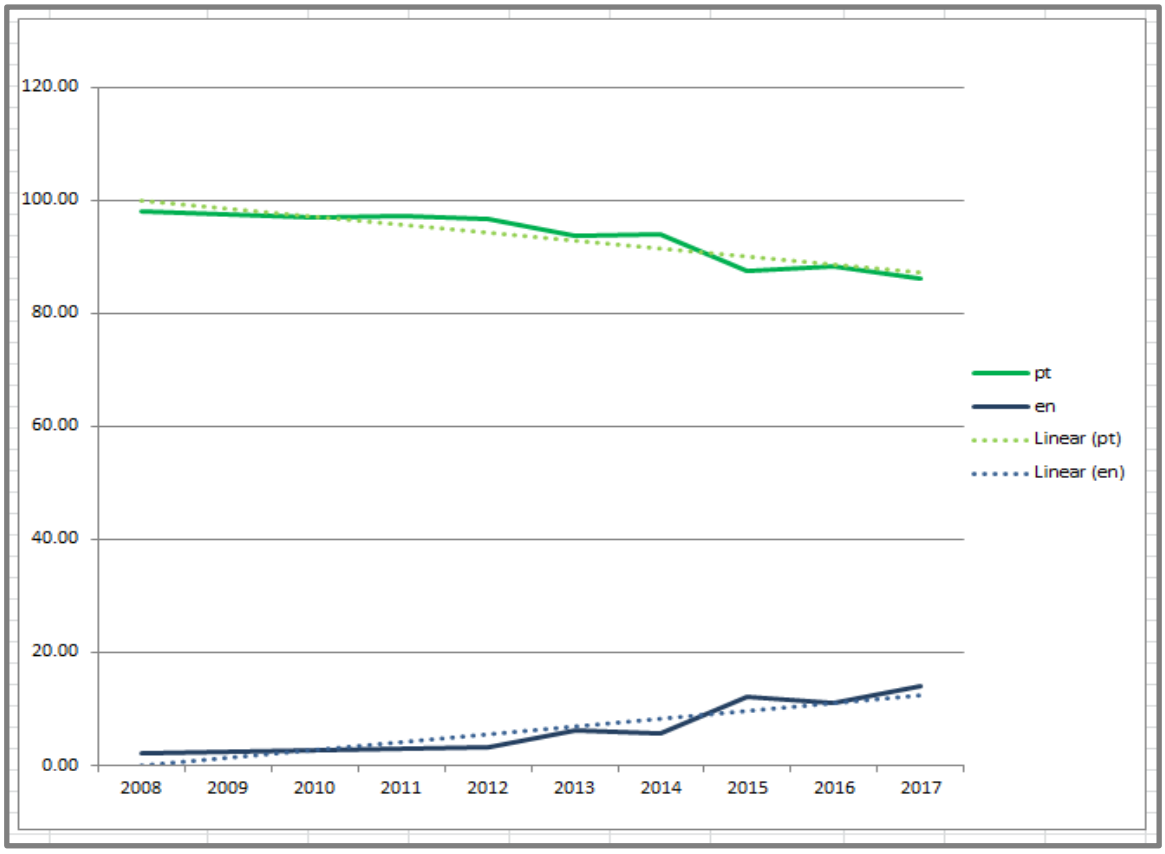

Figure 1 – Time series of the percentages of ETDs in Portuguese and in English from 2008 to 2017.

The percentages in figure 1 refer to all ETDs – Humanities & Theology, Science & Technology and Social Sciences. Analyzing raw data it is possible to see the differences among different graduate programs and areas. Figure 2 shows the numbers of graduate programs per percentages of ETDs in Portuguese.

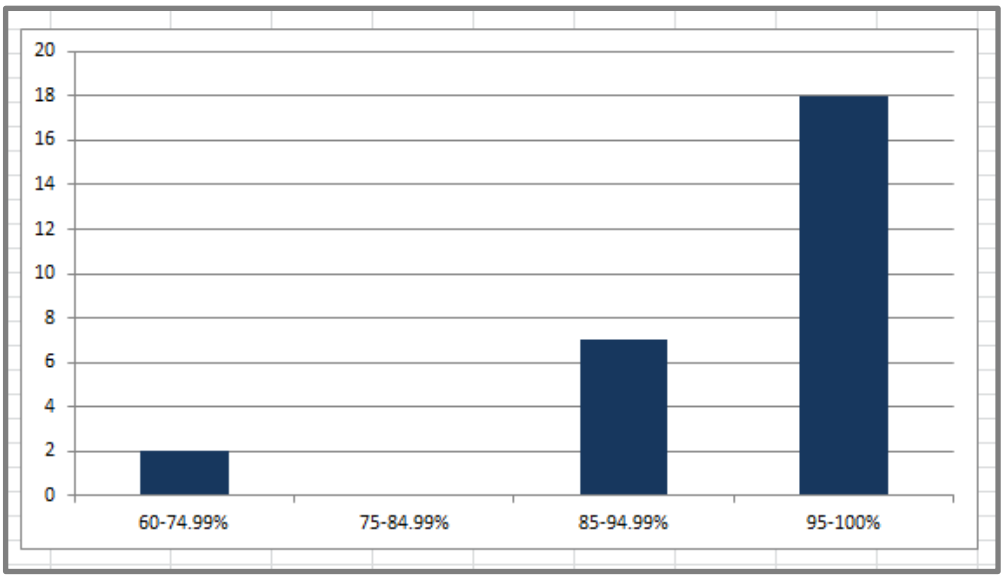

Figure 2 – Numbers of graduate programs per percentages of ETDs in Portuguese.

Some observations concerning figure 2 are necessary and three of them are based on the raw data (not shown in this work) that were examined:

- The total number of graduate programs in the figure is 27 while the number stated in section 1.b adds to 29. This happens because there are two sets of programs in the same areas and each pair was consolidated into one. One pair is in Humanities & Theology and the other in Social Sciences.
- The number of programs with 95% or over of ETDs in Portuguese is 18 (66.67% of all programs).
- Analysis of raw data showed that five programs had 100% of their ETDs in Portuguese.
- The programs with 95% or over are distributed as follows:
  - Humanities & Theology – 7 (100.0%)
  - Science & Technology – 4 (36.4%)
  - Social Sciences – 7 (77.8%)
- Only two programs have percentages lower than 75.0%. They are Economics (65.38%) and Informatics (71.40%). Economics belongs to the Social Sciences area and Informatics to Science & Technology.
- Graduate programs in Science & Technology provide 82.91% of all ETDs in English.
- ETDs in English are 11.26% of the ETDs in Science & Technology.
- The program of Chemistry has 0% of its ETDs in foreign languages and is quite an exception in the area of Science & Technology. It is the only graduate program in this situation.

The study addressed in this work started with the expectation that ETDs in English would have more accesses than the ones in Portuguese, which did not happen. Another expectation was that the percentage of accesses from Portuguese speaking countries would be lower than it is. The next section presents the methodology to compute accesses, the numbers and some interpretations.

03. ACCESSES TO ETDs IN DIFFERENT LANGUAGES @ PUC-Rio

Accesses to scholarly publications are the first steps to citations. Works that can not be found can not be read and cited. For this reason, it is important to examine if ETDs in English increase accesses; Spanish and French are not mentioned because the numbers are very low.

This section is devoted to examinig accesses to ETDs in English and Portuguese, and comparing numbers and countries. It is divided in three subsections – the first presents the way access data are collected and stored, the second addresses the criteria used to select the works to compare accesses and the third shows the numbers and interprets them.

a. Collecting Access Data

All contents on the system are described and have a sequential identification number. 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 on the system.

The Maxwell 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 runs at every xx:15 h and updates AWSTATS(http://awstats.sourceforge.net/) statistics displayed at https://www.maxwell.vrac.puc-rio.br/awstats/awstats.pl?config=www.maxwell.lambda.ele.puc-rio.br. The numbers shown in Open Access on this page are related to accesses to the system in general and not specifically to contents.

The second runs at every xx:01 h and updates the numbers of accesses to contents. The accesses are recorded in two tables on the database of the system and both have the same structure – columns for year, month, country code, content identification number and number of accesses. The difference between the tables is that one records all accesses while the other filters crawlers. All access data discussed in this work are store on the table with filtered numbers. The update works as follows:

- Every line of the log that is related to a content is identified by the content number and a digital format belonging to a specified set (pdf, doc, docx, xls, xlsx, ppt, pptx, mp3, mp4, wmv, htm, html). In case a new digital format is used, its extension is added to the set.
- Identified lines have the originating IP address matched to an international IP table that identifies the country to which the IP address is assigned. This table is automatically updated every week (at 2 AM on Mondays).
- The number of accesses to a content from a given country are added to the corresponding lines of the two tables on the database.

Access data started being collected in June 2004. So data are available for accesses in the last 14 years. It is not only data related to ETDs. There access statistics to all contents availale from the system and there is a complete set in public access tha can be viewed clicking Statistics on the left hand side menu of the homepage.

Contents that are partitioned, i.e., have more than on digital file will have the number of accesses equal to the sum of accesses to all files. This is the case of ETDs – when ETDs became mandatory the library decided that they should be partitioned with one file per chapter plus one file for cover/title page/abstract/etc and one for references; appendices and annexes should have separate file. This led to a situation that the average number of partitions was 7.3 per ETD. After this rule was dropped, the average number started decreasing and it currently is 6.1 This information is important because partitions have an impact on how accesses to different works are compared. A normalization was necessary for the comparison to be meaningful.

b. Selecting ETDs to be Examined

Section 02 presented the numbers of ETDs in foreign languages in different graduate programs.

The first important consequence of getting to know the numbers was not to consider works in Spanish and French. And the second was to decide the policy to select works to compare. It follows:

- The number of ETDs in English had to be higher than 10 in the program under consideration – this discarded many graduate programs that had less than 10 ETDs in English. (*)
- The percentage of ETDs in Portuguese should be smaller than 95%.(*)
- ETDs had to be public so accesses could come from users outside the PUC-Rio community.
- ETDs had to be published for, at least, 3 years so that indices in different seacrh engines and union catalogs were available.
- Comparisons were to be made among works in the same area and published approximately at the same time.
- The numbers of ETDs in Portuguese and in English for each area should be the same.
- The numbers of M and D works in each area should be the same if possible.

(*) These two conditions are not redundant since ETD collections of different graduate programs are quite diverse; programs do not have the same numbers of students.

These criteria led to the following graduate programs: Economics, Electrical Engineering, Industrial Engineering, Informatics , Mathematics and Mechanical Engineering. International Relations had a percentage that qualified (91.18%) but less than 10 ETDs in English. Among the six, the only graduate program not in the area of Science & Technology is Economics.

Since the Maxwell System offers many programs to query data on the database, the two that were chosen had different ways of showing retrieval results. The way data were showed was used not to introduce any bias. The works in English were selected based in alphabetical order of titles and the ones in Portuguese in alphabetical order of authors' names as long as all other conditions were met. They come from difeferent computer programs.

c. Data

When all criteria were applied to perform the searches, the numbers of ETDs that met them were:

- Economics – 14
- Electrical Engineering – 12
- Industrial Engineering – 04
- Informatics – 20
- Mathematics – 10

- Mechanical Engineering – 06

The total number is 66 – 33 in Portuguese and 36 in English. This number is 1.36% (66/4,866) of the ETDs published in and after 2008. Since the total number of ETDs in English is 275, 33 is 12.00% of this number.

Date found to match the criteria were quite different in terms of the time the ETDs had been published and the numbers of files the ETDs had. In order to compare data, it was necessary to define a normalized index. The option was to define index μ as:

$$\mu = \frac{\text{Number of acesses}}{(\text{Number of months})(\text{Number of files})}.$$

μ is the average number of accesses of an ETD per month and per file. It does not separate accesses by the language of the countries where they originate. The number of months (m) is computed from the month following the date of publication (the date is on the system) and the number of files (f) is on the system too.

Other numbers are examined and used to compare accesses. They are available from statistics programs on the Maxwell System in public access (https://www.maxwell.vrac.puc-rio.br/menu_estatistica.php?strSecao=etds) as well. They are:

- #C – number of countries that accessed an ETD.
- #accC – number of accesses from all countries.
- #ptC – number of Portuguese speaking countries that accessed an ETD.
- #accptC – number of accesses from Portuguese speaking countries.

They are used to compute other numbers. They are:

- μμ – average of the ETDs μ.
- μC – average number of countries.
- μptC – average number of Portuguese speaking countries.
- μm – average number of months after publication until July 2017.
- %accptC – percentage of accesses coming from Portuguese speaking countries.
- μ%accptC – average of percentages of accesses coming from Portuguese speaking countries.
- μ%accNptC – average of percentages of accesses coming from non Portuguese speaking countries.

Grad Prog	μm	μμ	μC	μptC	μ%accptC	μ%accNptC
Econ (en)	89.57	0.94	21.29	1.86	20.37	79.64
Econ (pt)	88.86	2.20	19.43	4.14	51.63	48.37
Elect Eng (en)	94.20	1.85	42.40	2.40	23.59	76.41
Elect Eng (pt)	92.60	5.43	39.80	5.20	73.97	26.03
Ind Eng (en)	91.50	2.55	67.50	2.00	21.66	78.34
Ind Eng (pt)	89.00	9.22	35.00	6.00	74.36	25.64
Inform (en)	90.30	1.41	41.10	1.90	21.07	78.93
Inform (pt)	90.90	3.30	26.20	3.70	39.72	60.28
Math (en)	94.00	1.82	43.60	1.60	16.02	83.98
Math (pt)	99.40	2.80	29.60	4.40	25.10	74.90
Mech Eng (en)	90.67	2.01	50.67	2.33	22.28	77.72
Mech Eng (pt)	87.67	4.96	33.67	4.33	69.91	30.09

Table 3 – Results of data compilation and processing for the variables previously defined (the last to columns when added yield 100%).

This table offers very interesting information that deserves to be commented:

- The average numbers of months after publication until July 2017 are in the narrow range (87.67 – 99.40) – this means that though ETDs had been published for different time frames, they were compatible in terms of order of magnitude. For ETDs in the same graduate programs the averages have narrower ranges.
- The average numbers of accesses for ETDs in Portuguese are higher than for ETDs in English for all graduate programs.
- The average numbers of countries that accessed ETDs in English are higher than the corresponding numbers for works in Portuguese for all graduate programs. Figure 3 clearly indicates this.
- The average numbers of Portuguese speaking countries that accessed ETDs in Portuguese are higher than the corresponding numbers for works in English for all graduate programs. Analysis of the countries that most accesses ETDs in English showed that they were Brazil and Portugal and in six cases only one country accessed. A comment on this is in the last section.
- The average percentages of accesses coming from Portuguese speaking countries are higher for ETDs in Portuguese than the corresponding numbers for works in English for all graduate programs. Figue 4 clearly indicates this.
- The average percentages of accesses coming from non Portuguese speaking countries are higher for ETDs in English than the corresponding numbers for works in Portuguese for all graduate programs. Figue 5 clearly indicates this.

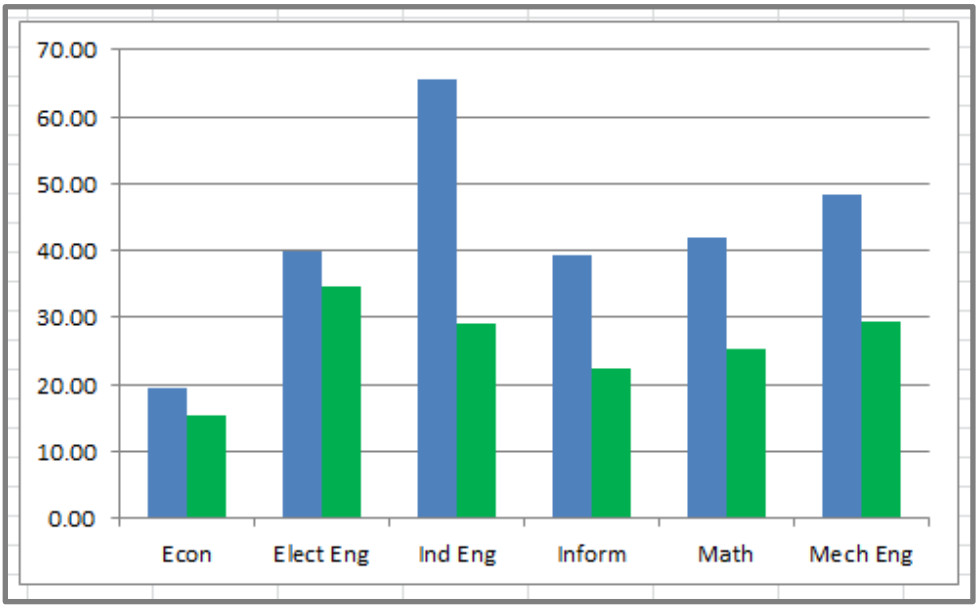

Figure 3 – Average numbers of countries that accessed ETD in Englsih (blue) and in Portuguese (green) by graduate program.

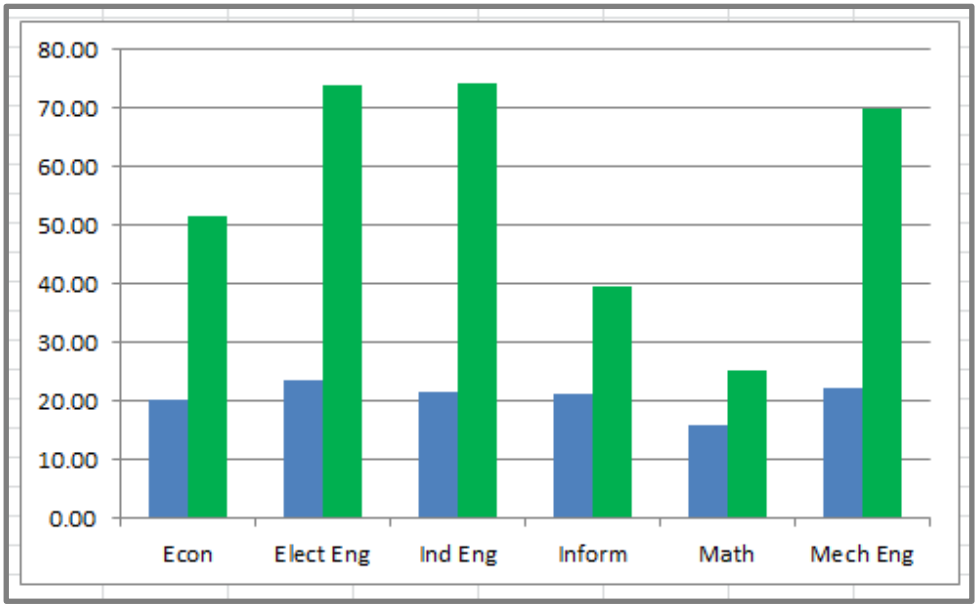

Figure 4 – Average percentages of accesses from Portuguese speaking countries to ETDs in Englsih (blue) and in Portuguese (green) by graduate program.

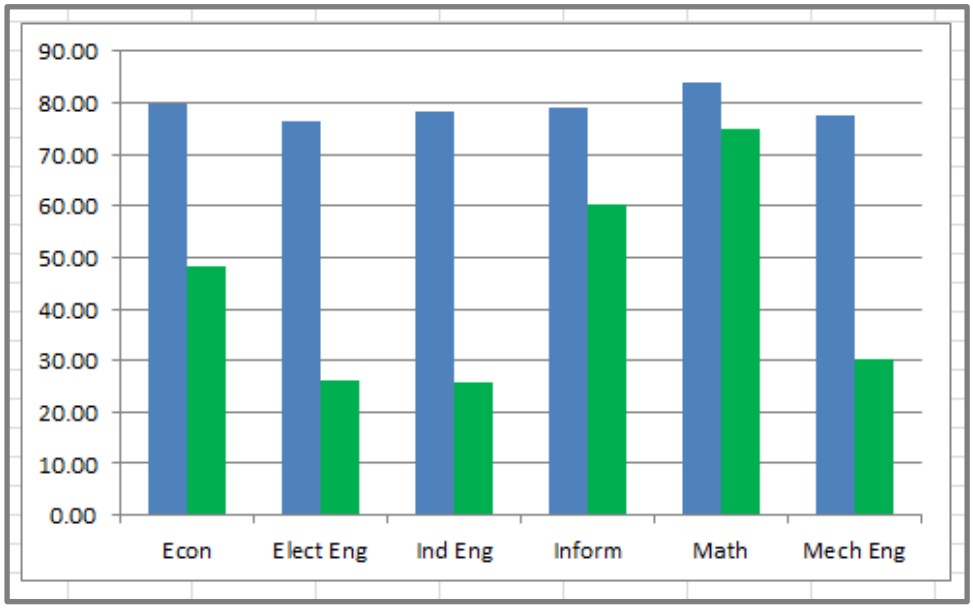

Figure 5 – Average percentages of accesses from non Portuguese speaking countries to ETDs in Englsih (blue) and in Portuguese (green) by graduate program.

Figures 4 and 5 show that the characteristics of accesses are completely different depending on country languages and the ETD languages – when they match and when they do not.

04. COMMENTS

Concerning accesses from Portuguese speaking countries to ETDs in English, it was mentioned in section 3 that in general they came from Brazil and Portugal. This is probably related to the lower educational level (as represented in the HDI and index I) of some countries – the number of persons who can read English must be lower in these countries than in Brazil (a much larger population) and in Portugal (a much higher HDI).

An interesting result is that examining numbers of accesses, ETDs in Portuguese have higher numbers (this can be seen in table 3). When it comes to the numbers of countries, ETDs in English have higher numbers. At the same time, the last two columns of the table (percentages of accesses) show that the percentages are higher when the languages match.

This work will be revisited in a year when more ETDs will satisfy the condition of being published for at three years. Some graduate programs did not have a considerable number of ETDs in English more than three years old; one example is Civil Engineering. A large sample may yield more accurate results and the permanent publishing of ETDs will make time series more consistent.

05. REFERENCES

[01] Simons, Gary F. and Charles D. Fennig (eds.). 2017. *Ethnologue: Languages of the World, Twentieth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com. Last accessed July 24, 2017.

[02] Central Intelligence Agency (CIA). 2017. *The World fact Book.* Washington DC, United States. Available: https://www.cia.gov/library/publications/the-world-factbook/fields/2119.html. Last accessed July 24, 2017.

[03] United Nations Development Programme (UNDP). 2016. *Human Development Index – International Human Development Indicators*. France. Available: http://hdr.undp.org/en/countries. Last accessed July 24, 2017.

[04] Pavani, Ana M. B. and Mazzeto, Ana C. E. 2010. Examining Accesses by Country and Language. Proceedings of ETD 2010 – 13th Symposium on Electronic These and Dissertations. United States. Available https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao+resultado&nrSeq=16848@2. Last accessed Julty 24, 2017.