The Lusophony Digital Humanities and what they (we) are doing from the South: textual corpus analysis and FAIR principles to tackle hegemony

Digital Humanities 2020, Ottawa.

Ricardo M. PIMENTA, IBICT, Brazil, orcid: https://orcid.org/0000-0002-1612-4126
Priscila R. CARVALHO, IBICT-UFRJ, Brazil, orcid: https://orcid.org/0000-0003-3042-1669
Josir C. GOMES, IBICT-UFRJ, Brazil, orcid: https://orcid.org/0000-0001-7629-1404

Introduction

The year 2018 witnessed a significant growth of research groups and laboratories dedicated to Digital Humanities in Brazil, however, without producing for this international community. Nowadays, the Digital Humanities are beginning to gain greater public interest in Brazil and other countries in South America.

In this perspective, our research would like to discuss what is produced in the Lusophone-speaking Digital Humanities in South America. The difficulty in diagnosing such production is evident because the global information regime has surrendered to the socio-technical and cultural monopoly mediated by Google, Amazon, Facebook, Apple, and Microsoft, major technological players (Fiormonte & Sordi 2019). In the case of the Portuguese-speaking world, it is noticeable the language barrier often puts the debate and dialogue-less in evidence. In addition, English literature as it is evidently English-speaking escapes from the larger question of problematization in the face of critical thinking, which is decolonization.

Methodology

The present work came from the research data of Gomes et. al. (2018) that retrieved Digital Humanities academic papers, thesis, and books written in the Portuguese language from Google Scholar. Despite Google Scholar's public access, it does not apparently provide consistent means that meet FAIR principles - Findable, Accessible, Interoperable, Reusable (Wilkinson et. al 2016), due to concentration and opacity of information retrieved, that it may be visible but not operable.

This empirical study analyzed 454 abstracts which composed the textual *corpus* through text mining techniques with IRaMuTeQ - Interface "R" for Multidimensional Analysis of Texts and Questionnaires (Marchand & Ratinaud 2012).

Results of IRaMuTeQ

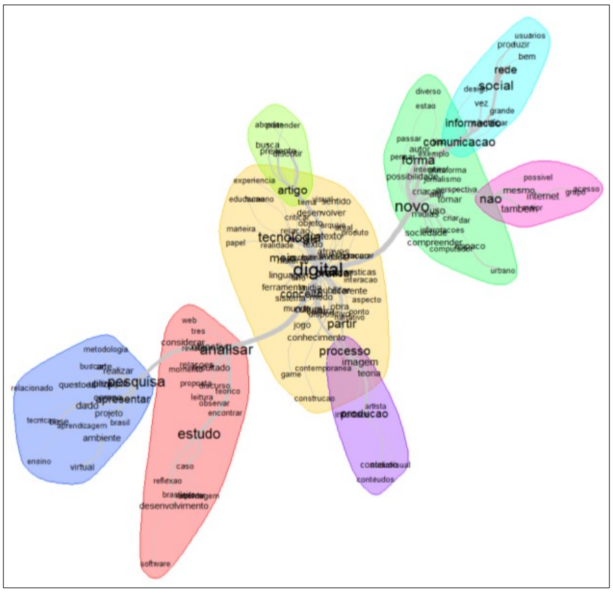

Figure 1: Similitude by IRaMuTeQ

The result of the similitude analysis, based on the graph theory, unveiled possible thematic convergences of the Portuguese language production in the Digital Humanities. The graph showed a central cluster represented by the term *digital,* which has a semantic attraction with the following terms: *conceito, texto, meio, ferramenta, linguagem,* and *interação*. In addition, this cluster has two subclusters identified by *artigo* and *processo*, as well it links other opposing clusters: *novo, estudo,* and *pesquisa.*

On the top right, the cluster *novo* has two subclusters, social and *nao*. In the first subcluster, the term *social* presented a possible connection between the highlighted terms: *comunicacao*, *informacao, social,* and *rede*. The second subcluster demonstrated a balance between the terms: *internet, possivel* and *acesso*. On the bottom left, the cluster *estudo* exposed the term *analisar* in evidence, and it links to the cluster *pesquisa* which is on the extremity of the graph.

From this analysis, it can infer the term set reflects actions, products, secondary research objects, and methods, besides the problems and challenges of non-internet access in South America.

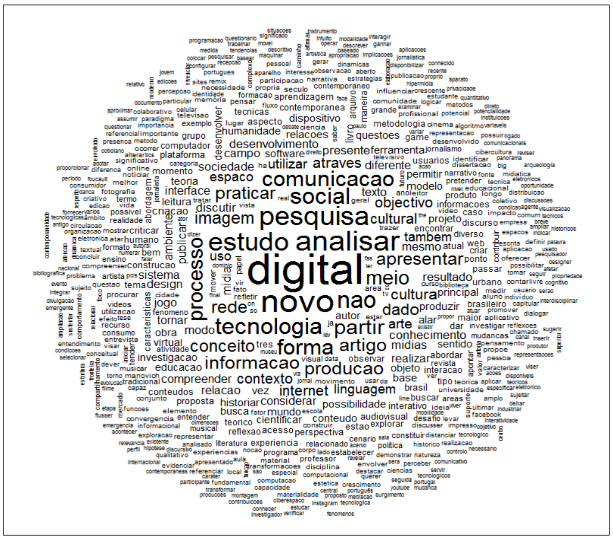

Figure 2: Word Cloud by IRaMuTeQ

The word cloud allowed the quick visualization and identification of the main keywords of the textual corpus: *digital, analisar, pesquisa, novo, comunicacao, pesquisa, tecnologia* and *nao.* Furthermore, this result reinforced the perception of the similitude analysis.

Conclusions

The usual bibliometric retrieval based on Web of Science and Scopus databases does not show the plethora of academic papers produced on Global South. From a decolonizing perspective, this study shows that scraping Google Scholar data could bring a broader result if you want to analyse Portuguese scientific production. In addition, the use of Zenodo allowed the research result to have a visibility to the public outside Brazil allowing that South America production could integrate Lusophone Digital Humanities in the global context, thus representing an important and necessary techno political action for researchers from that language community.

References

Fiormente, D., Sordi, P. (2019). "Digital Humanities of the South and GAFAM. For a Geopolitics of Digital Knowledge". Liinc em Revista, Rio de Janeiro, 15 (1), pp.108-130.

Gomes, J., Castro, F., Pimenta, R. (2018). "Google Scholar como fonte de medição da produção científica lusófona". LATmetrics - Anais do I Congresso de Altmetria e Ciência Aberta na América Latina. Dataset accessible: https://zenodo.org/record/1969123

Marchand, P., Ratinaud, P. (2012). "L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française". Actes des les Journées Internationales d'Analyse Statistique des Données Textuelles, pp.687-699.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data 3 (1): 160018. doi:10.1038/sdata.2016.18.