OPEN SCIENCE AND OPEN DATA TOWARDS OPEN EDUCATION — PORTLINGUE: A DIGITAL HUMANITIES RESEARCH PROJECT

Sílvia Araújo, Micaela Aguiar

Universidade do Minho (PORTUGAL)

Abstract

In this article, we will present the project PortLinguE — a project financed by European funds (ref. PTDC/LLT-LIG/31113/2017) and developed at the University of Minho. The aim of the project is to use open data to create resources for the scientific community and society in general and to empower the community itself to participate as content and knowledge creators. Three types of resources are being developed: (1) Resources for Specialised Languages: (a) a bilingual search engine (Portuguese-English), that presents a given term in both languages, in comparable texts, which are not the translation of each other; and (b) mappings of knowledge areas in primary, secondary and higher education, through the creation of glossaries and digital visualisations, using open access content, such as school programs, and through the direct collaboration with schools and students, encouraging the creation of learner-generated content and community participation in knowledge construction. (2) Resources for Academic Literacy: a seven-step method which intends to guide students in the preparation of academic assignments, by creating and aggregating of resources that support academic and scientific literacy, divided into seven steps that intend to guide students in the preparation of academic papers. (3) Resources for Science Communication: the dissemination of science is essential in a framework of open science and open education, so the project also focuses on promoting contact between the general population and researchers and scientists, by building sponsorship networks and scientific profiles of researchers.

Keywords: open science, open education, open data, research projects.

1 INTRODUCTION

The term “open” seems to be everywhere. Open Science, open education, open data, but also open access, open content or open source. The term seems to escape definition [1], however some defend that there is actually a strong consensus about what is meant by “open” [2]. “Open”, when it relates to education or educational technology, has two things in common: (1) “Free access to the content” and (2) “A formal grant of rights and permissions giving back to the user many of the rights and permissions copyright normally reserved exclusively for the creator or other rights holder” [2].

The concept of "open education" has been defined from different perspectives. For instance, open as in resources that don't have a cost for the user or open as in admission to formal education without entry requirements [3]. There are two major areas of development within open education: open educational resources and open educational practices. Open educational resources are “broadly defined as freely and openly accessible resources which are useful for educational purposes” [4]. Open educational practices move the focus from “content to practice and pedagogy”, by combining the use of open educational resources with “the goals of improving access, enhancing learning, and empowering learners” [3].

Open science can be defined as transparent and accessible knowledge that is shared and developed through collaborative networks [5]. The open science policy, which began to develop in the 2000s with the recommendations of the European University Association (EUA) in 2008 and materialised in Portugal with the Open Access Policy of the Foundation for Science and Technology (FCT) in 2014 and the Resolution of the Council of Ministers No. 21/2016, is based, according to the Open Science page, on the principles of Open Access, Open Data, Open Research and Innovation, Open Science Networks and Citizen Science.

Currently, Portugal has a consolidated knowledge management infrastructure, i.e., the Scientific Open Access Repository of Portugal [6], which provides access to over 700,000 documents from 300 national resources (165 scientific journals and 53 institutional repositories). National and European digital repositories have large amounts of text with immense potential to be exploited for the benefit of creating open educational resources. For example, journal articles and dissertations found on repositories make

up a valuable source of open data, from which scientific and academic terminology and phraseology can be extracted and reused to create new resources

The project PortLinguE sits at the intersection of open science, open data and open education. It aims to harness the potential of academic texts available in repositories to create a portal for open educational resources for specialised languages, academic literacy and science communication. The project is therefore oriented towards open education, from the perspective of open educational resources, as it uses open data to create new resources, and from the perspective of open educational practices, as we seek to "respect and empower learners as co-producers on their lifelong learning paths" [7] by promoting open learning architectures and citizen science. This paper provides an overview of the project and the resources being developed under its scope.

2 PORTLINGUE

PortLinguE is a Digital Humanities project financed by European funds, which emerges as an initiative of the Digital Humanities Group of the Centre for Humanistic Studies of the University of Minho. Due to its interdisciplinary nature, it is being developed by the School of Arts and Humanities (ELACH) and the School of Engineering (namely the Department of Informatics and the Department of Electronic Engineering) of the University of Minho, in a challenging and fruitful dialogue between areas such as Natural Language Processing, Machine Learning, Artificial Intelligence, Statistics, Corpus Linguistics and Lexicography.

This project aims to make new use of the abundance of open scientific data and open access texts on the web to create open educational resources for science creation and dissemination. Hence, we propose to build an online portal intended to provide its users (college students, professors, researchers, professionals, among others) with an array of open educational resources divided into three major sections: resources for Specialised Languages, resources for Academic Literacy and resources for Science Communication.

2.1 Resources for Specialised Languages

Modern society's tendency towards specialisation has shaped a new globalised and multilingual reality, where specialised languages are key to the commercial, political, social, and cultural worlds. The concept of specialised language (also known as special language, language for special purposes or language for specific purposes) generally refers to a language with specific features developed in response to the communicative needs of speakers, in a given area of expertise [8]. Indeed, specialised languages are the language of specialised knowledge ([9] defines specialised language as “a natural language considered as a vector of specialised knowledge”). As a result of open science policies, there is an abundance of open access data available on the web that offers an excellent opportunity to extract specialised information in Portuguese and other languages. Our goal is to make use of open access scientific data to create resources for specialised languages, which will be described in the following subsections.

2.1.1 Bilingual Search Engine

One of the resources is a bilingual search engine capable of identifying translation equivalents from comparable texts and of providing users with contextualised uses of languages for specific purposes in different domains of expertise and in different languages. The aim is to derive bilingual correspondences from the texts and thus to convert monolingual texts into bilingual terminology and phraseology. The innovation resides in the fact that the texts are not translations of each other. The focus will be, initially, to use Portuguese and English languages and to apply it to the scientific field of medicine. The search engine will feature a user-friendly query interface (similar to popular search engines, like Google) designed for simple queries without the use of regular expressions, which are typically complex and demotivating. Fig. 1 shows the summary of data flow.

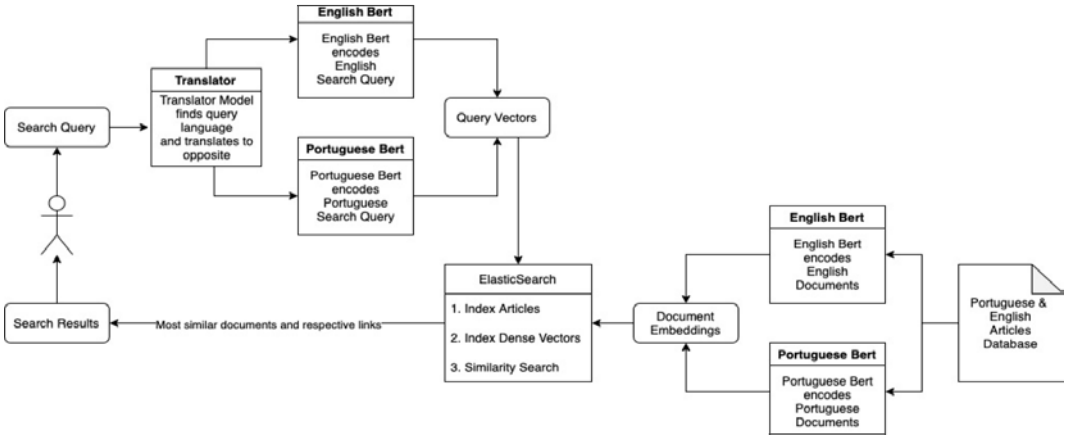

Figure 1. Summary of Data Flow

The search engine will work with BERT machine learning models [10]. BERT is a Natural Language Processing model that analyses text corpus in terms of similarities at the meaning level of words, collocations and sentences and distributes the processed data based on semantic similarity, thus generating semantic vectors. Two models are used, one pre-trained with a Portuguese corpus (Portuguese BERT) and the other with an English corpus (English BERT), which will transform users' queries and the articles extracted (by the Article Scraper) from scientific repositories (Google Scholar, Pubmed, RCAAP, for example) into semantic vectors. The articles as well as the respective semantic vectors are stored in an ElasticSearch database. After processing the queries and all the articles, semantic similarity mechanisms are applied. At the end of this process, the search engine will provide users with the most relevant results. As the vectors generated by BERT models are able to infer semantic context, the engine will be able to find relevant results even if the user has little to no knowledge of the specific vocabulary used in an area of expertise. We believe that our search engine will add value to the academic and professional environments, as the need for resources in specialised languages is becoming more and more common to an increasingly broader public.

2.1.2 Multimodal Glossaries and Visualisations

It is estimated that scientific knowledge doubles every five years [11], which results in a diversification of specialised languages [12]. Therefore, it is necessary to "inventory and register, as well as standardise the new lexical units" that arise from the "vertiginous scientific and technical progress" [12], as a way of valuing the language and safeguarding communication in specialised contexts. In this line, another one of our resources focuses on the creation of digital and multimodal glossaries and visualisations that map different areas of knowledge, using open access content, such as school programs, and through the direct collaboration with schools and students, encouraging the creation of learner-generated content and community participation in knowledge construction. Fig. 2 provides an example of a digital XML annotated glossary.

Figure 2. Digital XML Annotated Glossary

Creating glossaries involves researching typical expressions (terms of one or more words that usually appear together) and their respective definitions in a given area of knowledge in scientific articles. The glossaries' terms are annotated in XML format, according to the guidelines provided by the Text Encoding Initiative (TEI), to encode lexical resources (such as mono and multilingual dictionary and glossaries) readable by machines, mainly in the field of Social Sciences and Humanities. This flexible format allows the description of data in a consistent and structured way for electronic publications. The aim is to make available several versions of these lexicographical and terminological resources suitable for different target audiences and their needs. Mono-modal (text) glossaries will, for example, be practical for professionals, such as translators, interpreters and lexicographers, while transposing glossaries into a multimodal and interactive format will be more appealing to the general public and students, especially those in compulsory and higher education.

2.2 Resources for Academic Literacy

Academic literacy is rarely explicitly, systematically or comprehensively taught and with increasing levels of students pursuing postgraduate and doctoral studies [13], the lack of literacy in higher education poses a problem to open education. While some students do manage to master academic literacy skills, for most of them not having enough academic literacy constitutes an inequality factor and an accessibility issue, because without knowing how to communicate in academic contexts it is impossible to fully participate as a member of the academic community [14]. It is against this backdrop, that we are developing resources for academic literacy. Metodiza (Methodize, in English) is the name we have given to a resource that is both a seven-step method for academic assignments (oral presentation, scientific article, poster, dissertation, among others) and a platform of resources that are being created, structured around its seven steps. The intention is that each step of the method is accompanied by expository and informative materials in a variety of multimodal formats, such as videos and infographics, and useful digital tools that can be found in a single platform. Fig. 3 shows the schematization of the method in seven steps.

Figure 3. Seven-step method — Metodiza

The seven steps of the method are meant to guide students through the different stages of academic work, such as the planning and research stage (Prioritise, Search, Analise, Organise), the writing/creating stage (Textualise and Revise) and the dissemination stage (Finalise). This method aims to be applied by students on their own initiative or in pedagogical contexts, as part of an active methodology that aims to contribute to the autonomy and active participation of students in the construction of academic work.

2.3 Resources for Science Communication

It is estimated that half of the scientific articles produced worldwide will never be read by people other than their authors, editors and reviewers. Furthermore, 90% of articles never get cited. Scientific social networks [15], such as ResearchGate and Academia.edu, but also LinkedIn, Facebook, Twitter, are important in the context of open science and open education, as they begin to occupy a larger role in science communication and dissemination. In this way, we are also designing networking and sponsorship initiatives that put researchers and professionals in contact with the general public.

2.3.1 Lidera

Lidera is an initiative that intends to build a platform to bring together women that occupy high leadership positions in the field of technology and women who are just beginning their careers (senior undergraduates or recent graduates, for example). The participants would be able to tell their stories, share their learning, identify key skills to succeed and shed a light on relevant and current issues. The platform would enable a unique learning experience through active mentoring among some of the participating women, thus creating a sponsorship network.

2.3.2 Language Teaching Support Interface

The Language Teaching Support Interface is designed as a SVRE - Social Virtual Research Environment, which facilitates the sharing and management of information, encourages researchers to make their research objects available, in an environment that must be open and available for integration in other platforms. This initiative focuses on developing a multimodal guide that will help teachers to access text, video and audio information and support language teaching particularly in a context of social distance. Information relevant to language teaching will be made available in a digital graphic interface and interactive will be a tool at the service of open science. With the aim of facilitating and grouping in a single resource information relevant to language teaching, so that teachers can have access to materials of different natures (text, podcast and video). The goal is also to encourage researchers to record podcasts reporting on their research from discovery to approach, so that the practising community can have access to this knowledge and engage with them as well. The intention is to make research known in multimodal formats, and not just in the traditional text format.

3 CONCLUSIONS

The purpose of this paper was to provide a broad overview of the Digital Humanities research project PortLinguE, which works in the intersection of open science and open education principles, by making use of open scientific data to build new open educational resources for specialised languages, academic literacy and science communication.

ACKNOWLEDGEMENTS

This work was carried out within the scope of the “PortLinguE” project (PTDC / LLT-LIG / 31113/2017) financed by FEDER under Portugal 2020 and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT,I.P.).

REFERENCES

[1] C. Cronin. “Open education: Design and policy considerations” In R*ethinking pedagogy for a digital age: Principles and practices of design* (H. Beetham, & R. Sharpe eds.). London: Routledge, 2019.

[2] D. Wiley, “The Consensus Around ‘Open’”, *Improving learning*. Retrieved from: https://opencontent.org/blog/archives/4397

[3] C. Cronin, “Openness and Praxis: Exploring the Use of Open Educational Practices in Higher Education”, *The International Review of Research in Open and Distributed Learning*, vol. 18, no. 5, 2017.

[4] S. K. S. Cheung, K. C. Li, K.S. Yuen. “An Overview of Open Education Resources for Higher Education” in *Knowledge Sharing through Technology. ICT 2013. Communications in Computer and Information Science, vol 407* (J. Lam, K. C. Li, S. K. S. Cheung, F. L. Wang, eds). Berlin/Heidelberg: Springer, 2013.

[5] R. Vicente-Saez & C. Martinez-Fuentes, “Open Science now: A systematic literature review for an integrated definition”, *Journal of Business Research*, vol. 88(C), pp. 428-436, 2018.

[6] J. Carvalho, J. M. Moreira, E. Rodrigues, & R. Saraiva, “O repositório científico de acesso aberto de Portugal: Origem, evolução e desafios” in *Repositórios institucionais: Democratizando o acesso ao conhecimento* (M. J. Gomes, & F. Rosa eds.), pp. 127–152, EDUFBA, 2010.

[7] U. D. Ehlers, “Extending the Territory: From Open Educational Resources to Open Educational Practices”, *Journal of Open, Flexible, and Distance Learning*, vol. 15, no. 2, pp. 1-10, 2011.

[8] T. Afonso & S. Araújo, “Abordagem heurística das linguagens de especialidade com recurso à linguística de corpus: caso de estudo em linguagem jurídica”, *Polissema – Revista de Letras do ISCAP*, vol. 19, pp. 9–34, 2019.

[9] P. Lerat, *Les langues spécialisées*. Paris: Presses Universitaires de France, 1995.

[10] J. Devlin, M.-W. Chang, K. Lee, & K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *NAACL-HLT*, 2019.

[11] J. Cribb, & T. Sari, *Open Science: Sharing Knowledge in the Global Century*. Victoria: Collingwood, 2010.

[12] I. Gil, “Algumas considerações sobre Línguas de Especialidade e seus processos lexicogénicos”, *Máthesis*, vol.12, pp. 113-130, 2003.

[13] C. Ferrão Tavares, A. L. Pereira, “Apresentação: Literacias académicas multimodais.”, *Intercompreensão*, vol 16, pp. 5–10, 2012.

[14] A. M. Preto-Bay, “The Social-Cultural Dimension of Academic Literacy Development and the Explicit Teaching of Genres as Community Heuristics”, *The Reading Matrix* vol. 4, no, 3 pp. 86–117, 2004.

[15] J. Alonso-Arévalo, C. Lopes & M. Antunes, M. “Literacia da informação: da identidade digital à visibilidade científica” In *Literacia da Informação em Contexto Universitário* (C. Lopes, T. Sanches, I. Andrade, M. Antunes, & J. Alonso-Arévalo eds.), pp.109-152, Edições ISPA, 2016.