Linha D'Água: São Paulo, v. 34, n. 02, p. 1-8, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p1-8

Editorial

The Role of Linguistics in the Age of Digital Humanities

Rute Costa*
costamrv@gmail.com
https://orcid.org/0000-0002-3452-7228

Bruno Almeida**
brunoalmeida@fcsh.unl.pt
https://orcid.org/0000-0002-5777-5574

Margarida Ramos***
guida.ramos@sapo.pt
https://orcid.org/0000-0001-7209-3806

Maria Inês Batista Campos****
maricamp@usp.br
https://orcid.org/0000-0003-0004-9923

1 Brief considerations on digital humanities

In this issue 34/2 of *Linha D'Água*, we have chosen to prioritise the various fields of linguistics covering the analysis of different aspects, as well as the methodologies needed to comprehend, describe, formalise, and model them. The methodologies and theories that underpin the disciplinary areas predominantly addressed here – lexicography, terminology, text theory, and discourse analysis – bring together indispensable tools for a sustained analysis of lexicons and terminologies, texts and discourses, as well as of the knowledge produced in the subject areas that encompass the digital humanities, thus contributing to their development.

Since its inception in the 1940s, digital humanities have been understood as a research area that associates humanities and computing. However, according to Berry (2019), the disciplinary focus of the digital humanities has been expanding to include critical digital studies, as well as areas of knowledge more commonly associated with knowledge engineering, machine learning, data science and artificial intelligence. On the other hand, Piotrowski (2020, p. 3) draws attention to the fact that it is the adjective 'digital' in digital humanities that leads to varied interpretations, of which we retain three: (1) the use of digital tools and data; (2) the use of digital methodologies or methods; (3) research related to cultural phenomena and digital artefacts. In these three aspects, we find data, tools, methodologies, methods, and research that can

* PhD and researcher at Universidade NOVA de Lisboa, Lisbon, Portugal.
** PhD and researcher at Centro de Linguística da Universidade NOVA de Lisboa, Lisbon, Portugal.
*** PhD and researcher at Centro de Linguística da Universidade NOVA de Lisboa, Lisbon, Portugal.
**** PhD and researcher at Universidade de São Paulo, Brazil.

be applied to the humanities and to computing and technology since they interact with each other. While on the one hand, humanists of the digital age soon realised that computing and technology would play an increasingly central role in research in the humanities and social sciences (Berry, 2019), computing researchers, on the other hand, became aware of the need to gain access not only to data but also to the analytical tools specific to each disciplinary area under investigation. This could have an impact on and contribute to the development of new computational solutions more suited to the large quantity, diversity and complexity of the data that must be structured and shared.

Inter- and transdisciplinarity are at the very heart of digital humanities since the theoretical and methodological underpinnings of the different disciplinary areas intersect and contaminate each other, spawning a permanent dialogue between the various fields that make up the digital humanities. Data structuring requires a refined analysis of the data to be shared and reused according to Linked Data[1] principles, one of the biggest challenges the humanities and social sciences face.

The LMF[2], SKOS[3], TEI[4], and XML[5] standards and recommendations enable interoperability and data sharing at the intersection of humanities and the digital. By interoperability, we mean the ability of two or more systems or components to share information and use that information (Gerci, 1991, p. 42) in the most varied parts of the world.

Each of these standards has a specific function and can be used in a complementary way. ISO 24613-1:2019 'Language resource management – Lexical markup framework (LMF) – Part 1: Core model' proposes a common template to represent data found in mono- and multilingual lexical resources and thus enables their computational application. SKOS – Simple Knowledge Organization System – is a W3C recommendation to represent thesauri, classification schemes, taxonomies, authority lists, or any other type of controlled and/or structured vocabulary, whose primary purpose is to facilitate the publication and use of vocabularies as linked data. TEI, in turn, corresponds to a set of guidelines that specify coding methods for machine-readable texts, mainly in the humanities, social sciences, and linguistics. Finally, XML – Extensible Markup Language (XML) – is used to encode machine-readable texts. Currently, XML is also used in the exchange of data over the Web.

With the consolidation of the digital humanities, we are witnessing a paradigm shift in the disciplinary areas that integrate them due to the impact of the digital transition our societies are going through. This paradigm shift urges us to look at the disciplinary areas from a more

1 https://www.w3.org/standards/semanticweb/data

2 https://www.iso.org/standard/68516.html

3 https://www.w3.org/2004/02/skos/

4 https://tei-c.org/

5 https://www.w3.org/XML/

integrative perspective in order to enable them to interact better. In this context, linguistics needs to consolidate itself as a subject area, maintaining its identity so that it can sustainably bring added value to the digital humanities.

2 Linguistics and digital humanities

Linguistics, in its most diverse guises – terminology, lexicography, history of language, morphology, corpus linguistics, etc. – is a fully-fledged field of knowledge in the digital humanities. Linguistic resources and their underlying methodologies are objects of study in their own right, but they also provide support to other fields of knowledge – with their theoretical underpinnings, linguistic methodologies have traditionally been applied in furthering knowledge in the humanities, even before the digital age. Dictionaries and glossaries have always been designed to clarify or organise knowledge in the humanities, while text and discourse analysis has always been applied in the humanities. Some academic disciplines – e.g., archaeology, Egyptology, history, literature, information sciences, to name but a few – make use of linguistics in their research, theoretically and methodologically. Additionally, the fact that research in the humanities has become increasingly reliant on information technologies calls for a paradigm change, both regarding how it is conducted and how data are handled and made available to the international community.

In the digital age, as we experience it today, linguistic resources (terminological, lexical and textual) – viz., dictionaries, terminologies, glossaries, thesauri and controlled vocabularies, digital texts – represent a linguistic and cultural heritage, essential in a multilingual society. These resources occupy a crucial place in the digital humanities, whose field of study, which encompasses research as well as teaching, lies at the intersection between digital technologies and the various disciplines of the humanities.

On the other hand, the importance of methods and tools from computational linguistics, knowledge engineering, and text mining in applied research in the digital humanities highlights the relevance of linguistics for this field of study. These methods and tools imply recognising the value of corpus linguistics, which supposes a reflection on the criteria to constitute corpora and the application of linguistic knowledge to extract the information that needs to be analysed to serve the purposes of natural language processing.

3 About issue 34/2 of *Linha d'Água*

This issue comprises seven papers and a book review.

In the field of historical lexicography, **Geoffrey Clive Williams and Ioana Galleron** present a paper entitled *The hourglass effect: The late seventeenth encyclopaedic dictionary*

and the dissemination of knowledge, motivated by the retrodigitisation of Furetière's *Dictionnaire universel*, a work that marked the beginning of the encyclopaedic dictionary in 1690.

This retrodigitisation project aims to digitise various editions of the dictionary into the TEI[6] (Text Encoding Initiative) format, which is being carried out by adapting the GROBID-Dictionaries[7] software to historical dictionaries. Focus is then put on the knowledge-mediation role undertaken by the *Dictionnaire universel* by drawing on lexicographical and scholarly sources, a process the authors call the 'hourglass effect'. The analysis focuses on the 1701 edition, directed by Basnage de Beauval, in which the main compiler of scientific data, Regis of Amsterdam, used various botanical sources to write entries on the Brazilian flora. The conclusions highlight the role this work played in the phenomenon of the universal dictionary and the development of encyclopaedic works.

Bruno Almeida, in *Terminology and knowledge organisation: languages, vocabularies and systems*, proposes an analysis of the concepts underlying the terms 'documentary language', 'controlled vocabulary' and 'knowledge organisation system', based on the assumption that these tools can be understood as terminological resources.

The author envisions terminology as an interdiscipline through which multiple relationships between linguistics and the various fields of knowledge can be established. In this paper, the relationship with knowledge organisation, a sub-domain of information sciences, is explored using tools such as thesauri, classification schemes and other knowledge organisation systems. In this respect, SKOS (Simple Knowledge Organization System), a language used to represent knowledge organisation systems on the semantic Web, is assessed in terms of its ability to model terminological resources. The author's conclusions confirm the growing affinity between terminology and knowledge organisation, reflected in international standards and the applicability of SKOS to modelling terminological information.

Focusing on the documentation of endangered languages, in particular, Uralic languages, **Mika Hämäläinen**, **Jack Rueter** and **Khalid Alnajjar** describe in *Endangered language documentation in the digital age* an open infrastructure to build digital XML dictionaries with relevant applications for natural language processing (NLP).

The infrastructure described, Akusanat, is based on MediaWiki, which allows editing, searching, and visualising the contents of XML dictionaries. The solution described by the authors is used in the development of transducers, NLP tools that can lemmatise words, analyse their morphology and generate inflected forms. A Python library has also been developed to facilitate the use of dictionaries and transducers. The results allow Akusanat's infrastructure to

6 https://tei-c.org/

7 https://github.com/MedKhem/grobid-dictionaries

be interoperable with other NLP infrastructures dedicated to morphologically rich Uralic languages, such as Giella.

With the paper *Teaching Ancient Egyptian Language in Brazil: challenges and opportunities of digital resources*, **Ronaldo Guilherme Gurgel Pereira** and **Thais Rocha da Silva** present their project on Egyptian language teaching in Brazil using digital resources in the broader context of Egyptology education in the country.

The authors describe a case study on the Introduction to Classical Egyptian (Middle Egyptian) course designed in partnership with the ANPUH Ancient History Working Group (GTHA/ANPUH) and the Federal University of Santa Catarina (UFSC) that was taught between September and November 2020. This was the first course of its kind made available on a digital platform and in open access, making it possible for classes to run from Portugal to Brazil and Argentina. The results consolidated grammar as a work tool, combined with the availability of an anthology of sources and a public, digital and free access glossary. On the other hand, this experience aims to promote a collaborative environment among Brazilian Egyptologists, leading to the exchange of tools and digital resources and the consolidation of Egyptology in the country.

Still in the field of didactics, **Lukáš Zámečník** and **Ľudmila Lacková** propose a philosophical and methodological framework to teach digital humanities in universities, with the paper *Building Digital Humanities on the Linguistic Background: Methodological Basis for Digital Humanities Education in Undergraduate and Graduate Programs.*

Although digital humanities are often seen as a methodology or set of tools for data modelling, the authors advocate a broader perspective based on the confluence of the theoretical level in linguistics and the digital humanities tools. The paradigm endorsed by the authors through 'linguistic digital humanities' is based on the analysis of textual objects, qualitative linguistic concepts and, finally, the creation of new tools for data analysis and comparison. To exemplify the applicability of this paradigm in higher education, the paper presents two linguistics and digital humanities programmes at Palacký University in Olomouc, Czech Republic.

Looking at the phenomenon of deixis, **Miguel Magalhães** and **Matilde Gonçalves**, in *Deixis: A proposal for XML annotation within the text*, explore a methodology for the annotation of deictics in corpora to quantify these elements and visualise the construction of deixis in texts.

The literature review contextualises the annotation methodology on automatic text processing, deixis and XML structure and annotation. After this review, the authors explore applying the methodology in an analysis corpus comprised of texts selected in the scope of the activities of the Grammar & Text group of the Linguistics Research Centre of NOVA

University Lisbon (CLUNL). The criteria for the organisation of the corpus are canonicity, representativeness, and the language activity where the texts are included, namely journalistic, academic, or legal. The findings allow the quantification of spatial, temporal and personal deictic elements, as well as the establishment of relations between these elements and the language activity in which the texts are included. To conclude, the authors highlight the value of the proposed annotation for visualising the use of deictics in a text, enabling a better analysis. In the authors' words, the proposal has great potential for filling existing gaps and creating more flexible tools that can act at meso- and macro-textual levels and in less extensive but scalable corpora.

In the field of discourse analysis, **Ana Lúcia Tinoco Cabral** and **Manoel Francisco Guaranha** investigate the linguistic behaviour of social network users in *Digital interactions: conflict, argumentation, and verbal abuse on social media.*

The authors' research, contextualised in the interaction and construction of identities in social media, focuses on argumentation and polemics in social media platforms, particularly Facebook. A comment about COVID-19 vaccination in Brazil posted by a magazine on this platform prompted the authors to analyse user comments, focusing on polemics and identity. After analysing the data, the authors conclude that abuse revolves around the opposition between two poles – pro-vaccine / deniers – which corresponds to a verbal war voiced in those interactions and contributes to increasing disagreements between the parties involved in the digital environment.

Nathalia Akemi Sato Mitsunari presents a critical reading of Marie-Anne Paveu's work, *L'Analyse du Discours Numérique. Dictionnaire des formes et des pratiques*, a dictionary that has just been translated into Portuguese by Júlia Lourenço Costa and Roberto Leiser Baronas and published in 2017 by Editora Pontes.

This work, containing 31 headwords, describes concepts and categories for the analysis of digital discourse, or 'technodiscourse', also proposing an epistemological debate and citing studies on digital discourse in several countries, including Portugal and Brazil. The author of the dictionary assumes a cognitive position of discourse analysis, positioning herself in opposition to the French school of discourse analysis, namely its conceptions of context and interaction, which, according to Paveu, hinder the understanding of the specificity of native digital discourses. The 31 headwords that make up the dictionary reflect the theoretical position of the author.

Finally, we would like to thank the authors Geoffrey Clive Williams, Ioana Galleron, Bruno Almeida, Mika Hämäläinen, Jack Rueter, Khalid Alnajjar, Ronaldo Guilherme Gurgel Pereira, Thais Rocha da Silva, Lukáš Zámečník, Ľudmila Lacková, Miguel Magalhães, Matilde Gonçalves, Ana Lúcia Tinoco Cabral, Manoel Francisco Guaranha and Nathalia Akemi Sato

Mitsunari for having taken up the challenge launched by this issue 34/2 of the journal *Linha d'Água* focused on the interactions between research in linguistics and the digital humanities.

The articles published in this issue demonstrate the importance of linguistics in the digital humanities, as well as the multiplicity of perspectives and interdisciplinary approaches, including research in the field of lexicography (articles by Geoffrey Clive Williams and Ioana Galleron and by Mika Hämäläinen, Jack Rueter and Khalid Alnajjar), terminology (Bruno Almeida), the teaching of linguistics (Lukáš Zámečník and Ľudmila Lacková), text theory (Miguel Magalhães and Matilde Gonçalves), discourse analysis (Ana Lúcia Tinoco Cabral and Manoel Francisco Guaranha) and language didactics (Ronaldo Guilherme Gurgel Pereira and Thais Rocha da Silva).

Interestingly, although the authors' focus is linked to linguistic topics and areas, their backgrounds range from linguistics to other disciplinary areas, such as computer sciences, language technologies, Egyptology and philosophy, thereby doing justice to the digital humanities as an interdisciplinary research area.

The publication of this issue has received the support of the Programme of Support to Scientific Periodical Publications of the University of São Paulo/AGUIA, to which we are very thankful for allowing the indexation of *Linha d 'Água* on the Web of Science, a database of scientific papers produced by the Institute for Scientific Information and maintained by Clarivate Analytics, in the areas of Social Sciences, Arts and Humanities.

This journal counts on the work of the partners that make up its Editorial Board and *ad hoc* referees, and a body of Portuguese language top reviewers, all of whom guarantee its high quality. It also counts on the translation proofreading work carried out by Maria João Ferro, a researcher at the Linguistics Research Centre of NOVA University Lisbon (CLUNL).

With this issue of the journal, the Editorial Board seeks the internationalisation of the journal since we have received papers written by authors from several foreign universities, trying to meet the demands made by the University of São Paulo and the international agencies. The *Linha d'Água* journal has thus become an open space that gathers papers related to Portuguese language studies, linguistic-discursive studies and their relationship with teaching, maintaining a constant dialogue with the research developed in Brazil and abroad.

References

BERRY, D. M. What are the digital humanities?. *The British Academy*. Londres, 13 fev. 2019. Disponível em: https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities/. Acesso em 07 ago. 2021.

GERACI, A. *IEEE standard computer dictionary:* compilation of IEEE standard computer glossaries. IEEE Press, Piscataway, NJ, USA, 1991.

ISO 24613-1:2019 "Language resource management — Lexical markup framework (LMF) — Part 1: Core model", Genebra: ISO.

PIOTROWSKI, M. (2020, April 14). Ain't No Way Around It: Why We Need to Be Clear About What We Mean by "Digital Humanities". In: Wozu Digitale Geisteswissenschaften? Innovationen,Revisionen, Binnenkonflikt, 2020, Lüneburg, Anais, p. 1-16. DOI: https://doi.org/10.31235/osf.io/d2kb6. Acesso em: 07 ago. 2021.

São Paulo, August 2021.

Translated:
Maria João Ferro
Assistant Professor - Universidade NOVA de Lisboa, Portugal
mariajoaoferro@fcsh.unl.pt
https://orcid.org/0000-0001-8215-836X