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

# O papel da linguística na era das humanidades digitais

***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*******
mvramos@fcsh.unl.pt
https://orcid.org/0000-0001-7209-3806

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

## 1 Breves considerações sobre as Humanidades Digitais

Neste número 34/2 de *Linha D'Água,* optámos por dar primazia às mais diversas áreas disciplinares da linguística que abrangem a análise dos aspetos das línguas e da linguagem, bem como as metodologias necessárias para as compreender, descrever, formalizar e modelizar. As metodologias e as teorias que sustentam as áreas disciplinares aqui predominantemente em foco - lexicografia, terminologia, teoria do texto, e análise do discurso -, reúnem os instrumentos imprescindíveis para uma análise sustentada do léxico e das terminologias, dos textos e dos discursos, assim como dos conhecimentos produzidos nas disciplinas que congregam as humanidades digitais, contribuindo desta forma para o seu desenvolvimento.

Desde o seu surgimento, nos anos quarenta do século passado, que as humanidades digitais são entendidas como uma área de investigação que associa humanidades e computação. No entanto, de acordo com Berry (2019), o foco disciplinar das humanidades digitais tem vindo a alargar-se para incluir estudos digitais críticos, bem como áreas do saber mais comumente associadas à engenharia do conhecimento, aprendizagem de máquina, ciências de dados e inteligência artificial. Por sua vez, Piotrowski (2020, p. 3) chama a atenção para o facto de ser o adjetivo "digital" em humanidades digitais que leva a interpretações variadas, das quais retemos três: (1) o uso de ferramentas e dados digitais; (2) o uso de metodologias ou métodos digitais; (3) a investigação relacionada com fenómenos culturais e artefactos digitais. Nestes

* Doutora e pesquisadora da Universidade NOVA de Lisboa, Lisboa, Portugal.
** Doutor e pesquisador do Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal.
*** Doutora e pesquisadora do Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal.
**** Doutora e pesquisadora da Universidade de São Paulo, Brasil.

três pontos, encontramos os dados, as ferramentas, as metodologias, os métodos e a investigação que podem ser aplicadas às áreas das humanidades, mas também às áreas da computação e das tecnologias, uma vez que umas interagem com as outras. Se por um lado, os humanistas da era digital rapidamente tiveram a perceção de que a computação e a tecnologia teriam uma centralidade cada vez mais relevante na investigação das ciências humanas e sociais (Berry, 2019), os investigadores da computação, por seu turno, tomaram consciência da imprescindibilidade de ter acesso não só aos dados, mas também aos instrumentos de análise próprios a cada área disciplinar em estudo. Tal facto poderá ter um impacto e contribuir para a progressão da construção de novas soluções computacionais mais adaptadas à grande quantidade, diversidade e complexidade dos dados a estruturar e a partilhar.

A interdisciplinaridade e a transdisciplinaridade estão na essência das humanidades digitais, na medida em que os suportes teóricos e metodológicos próprios às diversas áreas disciplinares se entrecruzam e se contaminam, sendo requerido um diálogo permanente entre as várias disciplinas que perfazem as humanidades digitais. A estruturação dos dados exige uma análise apurada dos mesmos para serem partilhados e reutilizados de acordo com os princípios dos *Linked Data*[1], o que corresponde a um dos maiores desafios das ciências humanas e sociais.

Na intersecção das humanidades e do digital estão as normas e recomendações LMF[2], SKOS[3], TEI[4], XML[5] que permitem a interoperabilidade e a partilha de dados. Por interoperabilidade, entendemos a capacidade que dois ou mais sistemas ou componentes têm para partilhar informação e para usar a informação que foi partilhada (Gerci, 1991, p. 42) nos mais diversos pontos do mundo.

Cada uma destas normas tem uma função específica, podendo ser usadas de forma complementar. A norma ISO 24613-1:2019 "Language resource management — Lexical markup framework (LMF) — Part 1: Core model" propõe um modelo comum para a representação de dados presentes em recursos lexicais mono- e multilingues, de forma a permitir a sua aplicação computacional. O SKOS - Simple Knowledge Organization System - , é uma recomendação do W3C[6] para a representação de tesauros, esquemas de classificação, taxonomias, listas de autoridades, ou qualquer outro tipo de vocabulário controlado e/ou estruturado, sendo o seu principal objetivo facilitar a publicação e uso de vocabulários como *linked data*. Por sua vez, o TEI corresponde a um conjunto de diretrizes que especificam métodos de codificação para textos legíveis por máquina, principalmente nas humanidades, ciências sociais e linguística. Finalmente, a XML - Extensible Markup Language (XML) - é uma linguagem para a codificação de textos, de forma a serem legíveis pelas máquinas. Atualmente, a XML também é usada na troca de dados na Web.

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/

6 https://www.w3.org/

Com a consolidação das humanidades digitais, estamos a assistir a uma mudança de paradigma das áreas disciplinares que as integram e que resultam do impacto da transição digital pela qual estão a passar as nossas sociedades. Esta mudança de paradigma impele-nos a olhar para as áreas disciplinares de forma mais integradora, para permitir uma melhor interação entre as disciplinas. Nesta conjuntura, é importante a Linguística consolidar-se enquanto disciplina, mantendo a sua identidade para sustentadamente poder aportar mais-valia às humanidades digitais.

## 2 Linguística e humanidades digitais

A Linguística, nas suas mais variadas facetas – Terminologia, Lexicografia, História da Língua, Morfologia, Linguística de Corpora, etc. –, é uma disciplina de pleno direito nas humanidades digitais. Os recursos linguísticos e as metodologias subjacentes à sua conceção são objeto de estudo só por si, mas também são recursos de suporte em outras áreas do saber. As metodologias, com os respetivos suportes teóricos, são tradicionalmente aplicadas no aprofundamento do conhecimento nas humanidades, mesmo antes da era digital. Os dicionários e os glossários sempre foram pensados para esclarecer ou organizar conhecimento nas humanidades, enquanto a análise de textos e de discursos sempre foi aplicada nas humanidades. Disciplinas como a arqueologia, egiptologia, história, literatura, ciências da informação, só para enumerar algumas, recorrem à linguística nas suas investigações, tanto na vertente teórica como metodológica. Adicionalmente, o facto de a investigação em Humanidades se apoiar cada vez mais nas tecnologias da informação requer uma mudança de paradigma, tanto no seu estudo, como no tratamento dos dados e na sua disponibilização à comunidade internacional

Na era digital, tal como a vivenciamos hoje, os recursos linguísticos (terminológicos, lexicais e textuais) – dicionários, terminologias, glossários, tesauros e vocabulários controlados, textos digitais –, representam um património linguístico e cultural, essencial numa sociedade multilingue. Estes recursos ocupam um lugar central nas humanidades digitais, cujo domínio de estudo, que abarca a investigação, mas também o ensino, se posiciona na interseção entre as tecnologias digitais e as várias disciplinas das humanidades.

Por outro lado, a importância de métodos e ferramentas da Linguística Computacional, da Engenharia do Conhecimento e do *Text Mining* em investigação aplicada em humanidades digitais, evidencia a relevância da Linguística para esta área de estudos. Estes métodos e ferramentas implicam a valorização da linguística de corpora, que pressupõe uma reflexão sobre os critérios para constituir os *corpora* e a aplicação de conhecimento linguístico para a extração da informação que precisa de ser analisada para servir os intuitos do processamento natural da língua.

## 3 Acerca do número 34/2 da revista *Linha d'Água*

Este número é constituído por sete artigos e por uma resenha.

No âmbito da lexicografia histórica, **Geoffrey Clive Williams** e **Ioana Galleron** apresentam o artigo intitulado *O efeito da ampulheta: o dicionário enciclopédico do final do século XVII e a disseminação do conhecimento*, motivado pelo trabalho de retrodigitalização do *Dictionnaire universel* de Furetière, uma obra que marcou o início do dicionário enciclopédico em 1690.

O projeto de retrodigitalização da obra tem como objeto digitalizar diversas edições do dicionário para o formato TEI[7] (*Text Encoding Initiative*), um processo que está a ser levado a cabo através da adaptação do software *GROBID-Dictionaries*[8] a dicionários históricos. O enfoque é depois colocado no papel de mediação do conhecimento assumido pelo *Dictionnaire universel* através do recurso a fontes lexicográficas e eruditas, um processo que os autores apelidam de "efeito da ampulheta". A análise incide na edição de 1701, dirigida por Basnage de Beauval, na qual o principal compilador de dados científicos, Regis de Amesterdão, utilizou várias fontes botânicas para escrever entradas sobre a flora brasileira. As conclusões realçam o papel desta obra no fenómeno do dicionário universal e no desenvolvimento de obras enciclopédicas.

Por sua vez, **Bruno Almeida,** em *Terminologia e organização do conhecimento: linguagens, vocabulários e sistemas*, propõe uma análise dos conceitos subjacentes aos termos "linguagem documental", "vocabulário controlado" e "sistema de organização do conhecimento", partindo do pressuposto de que estas ferramentas podem ser entendidas como recursos terminológicos.

O autor posiciona a terminologia como interdisciplina, a qual permite estabelecer múltiplas relações entre a linguística e as diversas áreas do conhecimento. Neste artigo, é explorada a relação com a organização do conhecimento, um subdomínio da ciência da informação, por via de ferramentas como os tesauros, esquemas de classificações e outros sistemas de organização do conhecimento. Neste particular, o SKOS (*Simple Knowledge Organization System*), um modelo para a representação de sistemas de organização do conhecimento na *web* semântica, é avaliado em termos da sua capacidade de modelizar recursos terminológicos. As conclusões do autor vêm confirmar a crescente aproximação entre a terminologia e a organização do conhecimento, manifestada nas normas internacionais e na aplicabilidade do SKOS à modelização de informação terminológica.

Colocando o enfoque na documentação de línguas ameaçadas, em particular línguas urálicas, **Mika Hämäläinen**, **Jack Rueter** e **Khalid Alnajjar** descrevem, em *Documentación de lenguas amenazadas en la época digital*, uma infraestrutura aberta para a construção de

7 https://tei-c.org/

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

dicionários digitais em XML com aplicações relevantes ao nível do processamento de língua natural (PLN).

A infraestrutura descrita, *Akusanat*, baseia-se na *MediaWiki*, a qual permite editar, pesquisar e visualizar os conteúdos dos dicionários em XML. A solução descrita pelos autores é utilizada no desenvolvimento de transdutores, ferramentas do PLN que permitem lematizar palavras, analisar a sua morfologia e gerar formas conjugadas, tendo ainda sido desenvolvida uma livraria *Python* para facilitar o uso dos dicionários e transdutores. Os resultados permitem que a infraestrutura *Akusanat* seja interoperável com outras infraestruturas de PLN dedicadas às línguas urálicas, morfologicamente ricas, como é o caso da *Giella*.

Com o artigo *O ensino da língua egípcia clássica no Brasil: desafios e possibilidades usando recursos digitais*, **Ronaldo Guilherme Gurgel Pereira** e **Thais Rocha da Silva** apresentam-nos o seu projeto sobre a didática da língua egípcia no Brasil através de recursos digitais, no contexto mais vasto da formação em egiptologia no país.

Os autores utilizam como estudo de caso o curso *Introdução ao Egípcio Clássico (Egípcio Médio)* concebido em parceria com o Grupo de Trabalho de História Antiga da ANPUH (GTHA/ANPUH) e a Universidade Federal de Santa Catarina (UFSC), e ministrado entre setembro e novembro de 2020. Trata-se do primeiro curso do género disponibilizado em plataforma digital e em acesso aberto, possibilitando que as aulas fossem ministradas de Portugal para o Brasil e Argentina. Os resultados vieram consolidar a gramática como ferramenta de trabalho, aliada à disponibilização de uma antologia de fontes e glossário de acesso público, digital e gratuito. Por outro lado, esta experiência ambiciona promover um ambiente colaborativo entre os egiptólogos brasileiros, levando à partilha de ferramentas e recursos digitais e à consolidação da egiptologia no país.

Também no âmbito da didática, **Lukáš Zámečník** e **Ľudmila Lacková** propõem fundamentos filosóficos e metodológicos para o ensino das humanidades digitais nas universidades, com o artigo *Building Digital Humanities on the Linguistic Background: Methodological Basis for Digital Humanities Education in Gradual and Post-Gradual Programs*.

Embora as humanidades digitais sejam muitas vezes encaradas como metodologia, ou conjunto de ferramentas para modelar dados, os autores defendem uma perspetiva mais abrangente, baseada na confluência entre o plano teórico em linguística e as ferramentas das humanidades digitais. O paradigma defendido pelos autores através das "humanidades digitais linguísticas" tem como pilares a análise de objetos textuais, a utilização de conceitos qualitativos da linguística e, finalmente, a criação de novas ferramentas de análise e comparação dos dados. Para exemplificar a aplicabilidade deste paradigma no ensino superior, o artigo apresenta dois programas de linguística e humanidades digitais na Universidade Palacký em Olomouc, República Checa.

Olhando para o fenómeno da *deixis*, **Miguel Magalhães e Matilde Gonçalves**, em *A deixis: uma proposta de anotação em XML no âmbito do texto*, explora uma metodologia para a anotação de deíticos em *corpora*, permitindo quantificar estes elementos e visualizar a construção da *deixis* nos textos.

A metodologia de anotação é contextualizada pela revisão de literatura sobre tratamento automático de textos, *deixis* e estrutura e anotação em XML. Após esta revisão, os autores exploram a aplicação da metodologia num *corpus* de análise abrangendo textos selecionados no âmbito das atividades do grupo Gramática e Texto do Centro de Linguística da Universidade NOVA de Lisboa (NOVA CLUNL). Os critérios de organização do *corpus* consistem na canonicidade, na representatividade e na atividade da linguagem onde os textos se inserem, nomeadamente a jornalística, a académica e a jurídica. Os resultados permitem quantificar os elementos deíticos espaciais, temporais e pessoais, bem como estabelecer relações entre estes elementos e a atividade de linguagem em que se inserem os textos. Em conclusão, os autores destacam o valor que assume a anotação proposta para a visualização do uso dos deíticos num texto, possibilitando uma melhor análise. Nas palavras dos autores, a proposta apresenta um grande potencial para suprir falhas existentes e criar ferramentas mais flexíveis que possam atuar a níveis textuais meso e macro, e em *corpora* menos extensos, mas escaláveis.

No campo da análise do discurso, **Ana Lúcia Tinoco Cabral** e **Manoel Francisco Guaranha** investigam o comportamento linguístico dos utilizadores das redes sociais em *Interações digitais: conflito, argumentação e violência verbal nas redes sociais*.

A investigação dos autores, contextualizada na interação e construção de identidades nas redes sociais, incide sobre a argumentação e polémica nas redes sociais, em particular no *Facebook*. Um *post* de uma revista nesta rede social sobre a vacinação à COVID-19 no Brasil motiva os autores a desenvolver uma análise dos comentários dos utilizadores, com enfoque na polémica, na identidade. Após a análise de dados, os autores concluem que a violência gira em torno da oposição entre dois pólos - pró-vacina / negacionista -, que corresponde a uma guerra verbal manifestada nas interações e que contribui para aumentar as discordâncias entre as partes envolvidas em ambiente digital.

Por seu turno, **Nathalia Akemi Sato Mitsunari** apresenta-nos uma leitura crítica da obra de Marie-Anne Paveu, *L'Analyse du Discours Numérique. Dictionnaire des formes et des pratiques*, um dicionário que acaba de ser traduzido para o português por Júlia Lourenço Costa e Roberto Leiser Baronas, e publicado em 2017 pela editora Pontes.

Nesta obra, com 31 verbetes, são descritos conceitos e categorias para a análise do discurso digital, ou "tecnodiscurso", propondo ainda um debate epistemológico e citando estudos sobre o discurso digital em diversos países, incluindo Portugal e o Brasil. A autora do dicionário assume uma posição cognitivista da análise do discurso, posicionando-se em oposição à escola francesa da análise do discurso, nomeadamente às suas conceções de contexto e interação, as quais, segundo Paveu, colocam entraves à compreensão da especificidade dos

discursos digitais nativos. Os 31 verbetes que perfazem o dicionário refletem o posicionamento teórico da autora.

Para finalizar, gostaríamos de agradecer aos autores 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 e Nathalia Akemi Sato Mitsunari por terem respondido ao desafio lançado por este número 34/2 da revista *Linha d'Água*, incidindo nas interações entre a investigação em linguística e as humanidades digitais.

Os artigos publicados neste número vêm demonstrar a importância que a linguística assume nas humanidades digitais, assim como a multiplicidade de perspetivas e abordagens interdisciplinares, incluindo a investigação no âmbito da lexicografia (artigos de Geoffrey Clive Williams e Ioana Galleron e de Mika Hämäläinen, Jack Rueter e Khalid Alnajjar), da terminologia (Bruno Almeida), do ensino da linguística (Lukáš Zámečník e Ľudmila Lacková), da teoria do texto (Miguel Magalhães e Matilde Gonçalves), da análise do discurso (Ana Lúcia Tinoco Cabral e Manoel Francisco Guaranha) e da didática de línguas (Ronaldo Guilherme Gurgel Pereira e Thais Rocha da Silva).

Interessante é verificar que, apesar de o enfoque dos autores estar ligado a temáticas e a áreas linguísticas, os autores provêm, não só da linguística, mas também de áreas disciplinares tão diversas como as ciências da computação, as tecnologias da linguagem, a egiptologia e filosofia fazendo assim jus às humanidades digitais como área de investigação de natureza interdisciplinar.

A publicação deste número recebe o auxílio do Programa de Apoio às Publicações Científicas Periódicas da Universidade de São Paulo/SIBi, a quem agradecemos por permitir a indexação de *Linha d'Água* na Web of Science, base de dados de citações científicas do Institute for Scientific Information, mantida pela Clarivate Analytics, nas áreas de Ciências Sociais, Artes e Humanidades.

A revista conta com pareceristas do Conselho Editorial e *ad hoc* e com um corpo de revisores de língua portuguesa de excelência, o que garante sua alta qualidade. Conta também com o trabalho de revisão de tradução realizado por Maria João Ferro, investigadora do Centro de Linguística da Universidade NOVA de Lisboa.

Com este número da revista, o Conselho Editorial busca a internacionalização do periódico, uma vez que recebemos artigos de autores de universidades estrangeiras, procurando responder às exigências da Universidade de São Paulo e das agências internacionais. *Linha d'Água* torna-se, assim, um espaço aberto a publicações ligadas aos estudos de língua portuguesa, aos estudos linguístico-discursivos e sua relação com o ensino, mantendo um diálogo constante com as pesquisas desenvolvidas no Brasil e no exterior.

## Referências

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, agosto de 2021.