V. 34 n.2

ISSN 2236-4242

mai-ago 2021

LINHA D'ÁGUA

Programa de Pós-Graduação em Filologia e Língua Portuguesa
Faculdade de Filosofia, Letras e Ciências Humanas
Universidade de São Paulo

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.

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

Linha D'Água: São Paulo, v. 34, n. 02, p. 9-25, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p9-25

Artigo / Article

The Hourglass Effect: The Late Seventeenth Encyclopaedic Dictionary and the Dissemination of Knowledge

O efeito da ampulheta: o dicionário enciclopédico do final do século XVII e a disseminação do conhecimento

Geoffrey Clive Williams*
williams@licorn-research.fr
https://orcid.org/0000-0001-8790-7534

Ioana Galleron**
ioana.galleron@sorbonne-nouvelle.fr
https://orcid.org/0000-0003-0393-4485

Abstract

The publication of the Dictionnaire universel of Furetière in 1690 ushered in the age of the encyclopaedic dictionary. This was a relatively short-lived phenomenon of little more than a hundred years, but one which pathed the way to modern encyclopaedias. Furetière having died in 1688, his successor was Basnage de Beauval, a protestant exile based in the United Provinces of the Netherlands. It was Basnage who in the new 1701 edition transformed the dictionary by enlarging it considerably to a more genuine encyclopaedic coverage and calling on specialists to rewrite key sections, notably on the natural sciences. The simile of the hourglass is a means to show how the dictionary mediated knowledge from a vast array of sources and made the data available to contemporary and current day users. This paper demonstrates the hourglass effect through the lexicographical and learned sources that Basnage and his major compiler of scientific data, Regis of Amsterdam, brought into service. It looks at how Regis used numerous botanical sources in writing entries on Brazilian flora. Finally, we examine the influence of the work on the phenomenon of the universal dictionary and the development of the encyclopaedia.

Keywords: Historical Lexicography; Retrodigitisation; Encyclopaedic Dictionaries; Source Tracking; Botany.

* Université Grenoble Alpes & Université Bretagne Sud, UMR 5316 Litt & Arts – UGA, Grenoble, France.
** Université Sorbonne-Nouvelle Paris 3, UMR 8094 LATTICE – USN, Paris, France.

Resumo

A publicação do Dictionnaire universel de Furetière em 1690 marcou o início da era do dicionário enciclopédico. Este foi um fenómeno relativamente curto com pouco mais de cem anos, mas que abriu caminho para as enciclopédias modernas. Tendo Furetière falecido em 1688, seu sucessor foi Basnage de Beauval, um protestante exilado sediado nas Províncias Unidas da Holanda. Foi Basnage quem, na nova edição de 1701, transformou o dicionário, ampliando-o consideravelmente para uma cobertura enciclopédica mais genuína e chamando especialistas para reescrever secções-chave, principalmente sobre as ciências naturais. A analogia da ampulheta é uma maneira de mostrar como o dicionário mediou o conhecimento de uma vasta gama de fontes e disponibilizou os dados para utilizadores contemporâneos e atuais. Este artigo demonstra o efeito de ampulheta através de fontes lexicográficas e eruditas que Basnage e seu principal compilador de dados científicos, Regis de Amsterdão, colocaram em serviço. O artigo examina como Regis utilizou várias fontes botânicas para escrever entradas sobre a flora brasileira. Por fim, examina a influência da obra no fenómeno do dicionário universal e no desenvolvimento da enciclopédia.

Palavras-chave: *Lexicografia histórica; Retrodigitalização; Dicionários enciclopédicos; Rastreio de origem; Botânica.*

Introduction

The publication of the *Dictionnaire universel* of Furetière in 1690 ushered in the age of the encyclopaedic dictionary. This was a relatively short-lived phenomenon of little more than a hundred years, but one which pathed the way to modern encyclopaedias. Furetière having died in 1688, his successor was Basnage de Beauval, a protestant exile based in the United Provinces of the Netherlands. It is Basnage who, in the new 1701 edition, transformed the dictionary by enlarging it considerably to a more genuine encyclopaedic coverage and calling on specialists to rewrite key sections, notably on the natural sciences. The simile of the hourglass is a means to show how the dictionary mediated knowledge from a vast array of sources and made the data available to contemporary and current day users. This paper demonstrates the hourglass effect through the lexicographical and learned sources that Basnage and his major compiler of scientific data, Regis of Amsterdam, brought into service. More specifically, it looks at how Regis used numerous botanical sources in writing entries on Brazilian flora. Finally, it discusses the influence of the work on the phenomenon of the universal dictionary, and the development of encyclopaedia.

Dictionaries are often seen as the supreme reference and arbitrator of word meaning. The notion that a dictionary says what a word really means is fundamentally wrong. This is a frequent and unfortunate fallacy. The history of lexicographical endeavours, from glosses via catholicons to dictionaries, is one of increasing decontextualization, when meaning can only really be found in context. Lexicographical elegance (RUNDELL, 2010) is about concision, reducing language knowledge to a minimum. However, what that means is that rather than being a concentration of knowledge, dictionaries are essentially tools for disambiguation. They

represent the knowledge of the lexicographer, or the analysis made by a lexicographer from a large amount of data, and nowadays this means electronic corpus. As (HANKS, 1986) has so nicely put it, a dictionary can at best demonstrate meaning potentials, not meanings. Dictionaries, as opposed to other works of reference, essentially deal with words rather than *things*. On the other hand, encyclopaedias do attempt to encapsulate knowledge, and provide means to dig further into a subject. They are essentially about *things* and have developed since their beginnings, in works as Chambers' *Cyclopaedia* and the great encyclopaedia of Diderot and d'Alembert, to become vast repositories of knowledge, typified nowadays by Wikipedia.

However, the difference between dictionaries and encyclopaedias has not always been so clear cut. In the late seventeenth century there arose a new type of dictionary, one that dealt with both words and things, and which endeavoured the be a reference work for knowledge. Its greatest exponent is probably the *Dictionnaire Universel*, henceforth DU, of Antoine Furetière (FURETIÈRE, 1690) and most particularly the 1701 edition under the direction of Henri Basnage de Beauval (FURETIÈRE, 1701). It is this work that is the direct ancestor of the modern encyclopaedia and it principal author, abbé Antoine Furetière (1619-1688) was, in the words of Rey (REY, 2006), a precursor of the age of enlightenment. Furetière had been unable to publish his work in France, and so through the good offices of the protestant exile Pierre Bayle, it was published in the United Provinces of the Netherlands.

In this paper, we shall describe the ongoing work on digitising the DU (WILLIAMS; GALLERON, 2016) and insight the management of some the data through what we call the hourglass effect. The hourglass effect attempts to illustrate how large amounts of data are compiled as input mediated through definitions that then act as a knowledge base for users. Input here takes the form of earlier dictionary sources and the array of other works consulted by the compilers. The range of texts that may be used in encyclopaedic entries is illustrated through the botanical hourglass seen in the context of descriptions of Brazilian trees. We then look at how idea of a universal dictionary was taken up and how the DU influenced, directly or indirectly, other works.

Digitising the DU

In carrying out the analyses in this paper, we make use of a fully digitised version in XML-TEI format. Whilst a first attempt at digitising the DU is described in (WIONET; TUTIN, 2001), this project worked on a 1702 edition, which seems in fact to be a two-volume pirate version printed in France, and simply copying the true three-volume 1701 version. Only the letter D was attempted, and the files were never made available. Thus, more recently in 2015 an independent project was started within the context of the Consortium Cahier[1] of HumaNum, the French national scheme for the development of digital humanities. This initial project

[1] Available at: https://cahier.hypotheses.org Last accessed: 03 mar. 2021.

consisted in working from scratch on the letter C in XML-TEI as a means to explore the possibilities of in-depth mark-up and to prepare the ground for full digitalisation of the entire work. Full digitalisation became possible with French National funding for a consortium built from research teams in the universities of Grenoble-Alpes and Sorbonne-Nouvelle, in collaboration with Inria, the French national institute for applied computing[2]. The aim is to digitise all four major editions – 1690, 1701, 1708 and 1725/27 –, as well as the rival Trévoux dictionary of 1704, in fact a total plagiarism of the 1701 DU carried out by Jesuits in an attempt to undermine the protestant Dutch publishing industry.

The digitalisation process is a massive one, in that the work consists of over 3000 pages of text organised in two columns. This is far too much for human intervention, so the task of preparing a TEI version was carried out using GROBID-Dictionaries (KHEMAKHEM *et al*., 2017), by adapting this software to the needs and particularities of historical dictionaries. GROBID-Dictionaries requires an OCR in a dual layer PDF so as to have both the initial layout and the text. A high quality OCR is essential (KHEMAKHEM *et al*., 2019), so we had the volumes rescanned by the *Bibliothèque Nationale de France* in order to have a higher degree of resolution before undertaking an OCR with Transkribus[3]. Numerous problems had to be overcome, due to page layout and the use of characters that are no longer in current use, such as the s-long which OCR frequently interprets as an *f*. Nevertheless, a usable working edition was obtained that was suitable for parsing by GROBID-Dictionnaries, and for use in a concordancer with a full private version being uploaded to SketchEngine.

GROBID-Dictionnaries works by tackling each level of analysis separately. This entails training the machine learning system at each level. The tool has performed extremely well down to the level of sense and sub-sense so that we now have a perfectly usable full version. Nevertheless, correcting at sense level is an ongoing and time-consuming process, as this is where a less than perfect OCR and the idiosyncrasies of a late seventeenth century dictionary make machine mark-up difficult. Once the sense level is correct, a first stage of crowd-sourced correction will take place using the TACT system developed at the University Grenoble Alpes[4].

In addition to structural mark-up with Grobid dictionaries, work is underway to apply a Named Entity Recognition system using BERT technology under the designation CamemBERT (ORTIZ SUAREZ *et al*., 2020). Apart from the standard targets of NER, the challenge is to specifically locate the persons and works cited in the text. This is complicated by the need to differentiate persons who are mentioned, such a Julius Caesar and Jesus Christ, from persons who are actually cited, and from the use of abbreviations used for persons or texts when citing sources. Basnage supplied a list of persons cited with the abbreviations in the front matter of his edition. Unfortunately, not only the abbreviations are not always respected, but

2 Available at: https://anr.fr/Projet-ANR-18-CE38-0003 Last accessed: 03 mar. 2021.

3 Available at : https://transkribus.eu/lite/ Last accessed: 03 mar. 2021.

4 Available at : https://tact.demarre-shs.fr/actu Last accessed: 03 mar. 2021.

also numerous variants are to be found. In addition, a very large number of other authors and works are cited. It is hoped that NER system will ultimately help with automatic mark-up of citations, as this was done manually in the work on letter C.

In the meanwhile, a prosopographical profile is being built and stored in the TEI header for persons cited. Some of these, with abbreviations that often require a fair degree of disambiguation, come from an index provided by Basnage. It is not known how he compiled this list as far more works are actually cited. To date, 240 authors have been found and linked to a profile. We are fortunate that a detailed study has already been made of the life of Basnage (GRAVELEAU, 2018), so we can use this as a starting point. Dr Graveleau is currently within the project, working building detailed profiles of each author and also listing their works and the availability of those works as digital editions. In this way, we hope to obtain a deeper insight into the sources and construction of the entries. We wish to know who and what influenced the dictionary as well as mapping the influence it has in turn on other lexicographers and encyclopaedists.

The hourglass effect

The simile of hourglass can be a useful one in describing dictionaries insofar as these are essentially tools for mediating linguistic and encyclopaedic information. A good dictionary is based on the compilation of vast amounts of knowledge that has to be passed through the filter of lexicographical elegance (RUNDELL, 2010), that is the use of carefully constructed concise entries, before being gradually dispersed and appropriated by users. Past dictionaries often tended to have a literary bias and have citations, if any, drawn from so-called *best authors*. This is why the Dictionnaire *universel* of Furetière is so revolutionary in that the work, particularly in its 1701 edition by Basnage de Beauval, drew upon the most appropriate authors. Rather than prestige, the citations seek to show both usage and knowledge from the best and most up-to date sources. It is essentially a precursor of the age of the encyclopaedia (ROY-GARIBAL, 2006). As such, the dictionary supplied contemporary readers with much valuable information and offers us an interesting insight into arts and sciences of the 17th century, all the more so as the sources cited give the possibility of creating a virtual library, a corpus in many ways, of the works and authors consulted.

The prefaces to the dictionary give interesting insights into the motivations and methods (WILLIAMS *et al*., 2020). In the preface to the 1690 edition, written by the French philosopher Pierre Bayle as Furetière had died in 1688, the encyclopaedic nature is underlined:

> On ne sera plus reduit, comme le sont tant de gens, dans les matieres même les plus communes, à recourir au mot vague de chose, de piece, & à faire des postures de mains & de pieds, (manieres qui passent avec raison pour rustiques) afin d'exprimer la figure, la situation, & l'étenduë de ce dont on parle. Cet Auteur apprend à tout le monde, non seulement la nature des choses par leur matiere, leurs

> usages, leurs especes, leurs figures, & leurs autres proprietez, mais aussi les termes propres dont il se faut servir pour les décrire.[5]
>
> (One will not be obliged any more to use the vague words of 'thing', 'piece', as most of people do, even when speaking about very common things, or to gesticulate with hands and feet (all manners that are rightly considered as rustic) so as to indicate the figure, the position or the size of what one is talking about. This author teaches everybody the nature of each thing, its materials, its usages, the different sorts it comes in, but also the appropriate terms for describing these.) (BAYLE, 1690, s. p.)

This explains the wealth of terms in the dictionary, but also some of the inaccuracies in that although he consulted many sources, the work was essentially that of Furetière. What is most significant with Basnage, is that he recognised his own lack of knowledge in key areas and so called upon experts, notably a certain Mr Regis of Amsterdam:

> Je ne mets pourtant pas sur mon compte les articles d'Algèbre. Cette science m'est inconnue. Je ne m'approprie point non plus ce qui regarde la Medecine, l'Anatomie, la Pharmacie, la Chirurgie, & la Botanique. Je n'ai point voulu me fier à moi-même là-dessus. Un habile Mr. Regis, Medecin à Amsterdam homme s'en est chargé (FURETIÈRE, 1701, s. p.)[6]

It is precisely this use of experts and expert sources that makes the dictionary so significant and makes the hourglass effect so visible. The DU has many facets so a diversity of hourglasses can be observed, here we deal with two: lexicographical sources and botanical information.

[5] Please note that all citations are in the French of the period. The dictionary has no page numbers, in any of its editions, therefore only the main sections ("Préface", tome) are indicated with the quotations.

[6] In English: "I do not declare as mine the entries about algebra. I do not know this science. I do not claim either anything pertaining to medicine, anatomy, pharmacy, surgery and botanical sciences. I did not want to trust myself on these matters. A knowledgeable man, Mr. Regis, medical practitioner in Amsterdam, took charge of these".

The lexicographical hourglass

Fig. 1: 16th century hourglass[7]

Furetière was at great pains to show that the dictionary was his own work and was in no way influenced by that of the Academy. The attempts by the Academy on what they saw as a rival dictionary, by preventing publication through use of their 40 years monopoly on producing a dictionary for the French language, led to open warfare and ultimately to Furetière's exclusion from the Academy. His bitterness was made quite clear in his *Factums*, the publications he wrote in his defence. Basnage did not have this problem. As an exiled Protestant, he was even keen to show his links with France and was therefore happy to compile his dictionary by making extensive use of published works, including that of the Academy and the *Dictionary of arts and sciences* of Corneille.

[7] Metropolitan Museum of Art, CC0, via Wikimedia Commons. Available at: https://commons.wikimedia.org/wiki/File:Half-hour_sand_glass_MET_ES268.jpg Last accessed: 03 mar. 2021.

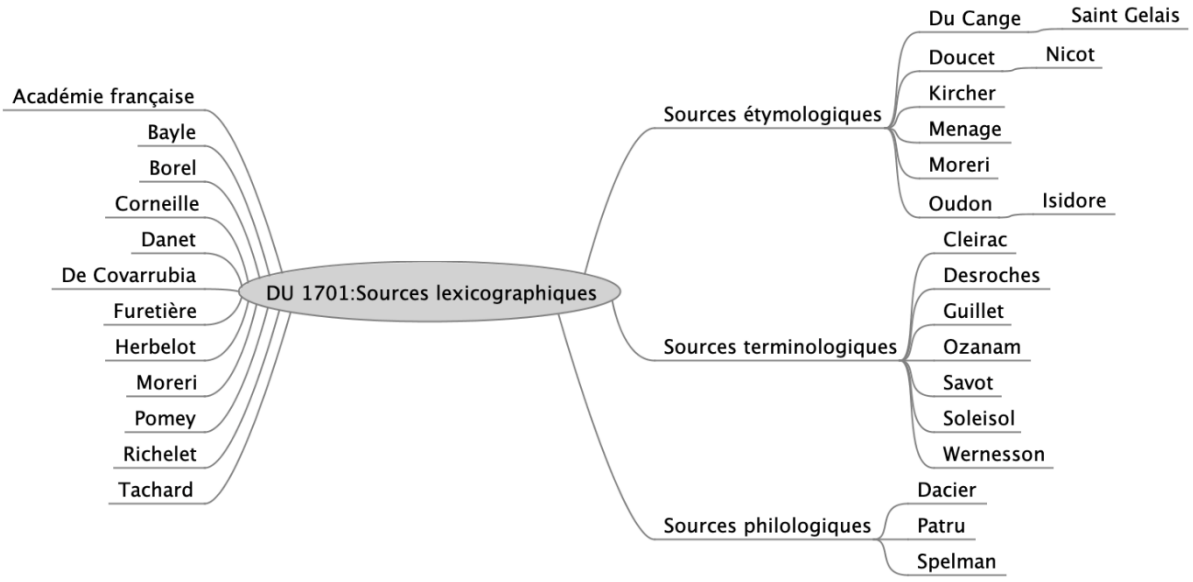
Basnage cites other lexicographical works essentially when giving word histories and this is well illustrated in the entry for the noun *Halte* which cites the grammarian Vaugelas as well as the dictionaries of the Academy and Richelet.

> HALTE. subst. f. Le h s'aspire. Terme de Guerre, […] D'autres disent que c'est un mot Allemand, car on dit halten en même signification. Selon Vaugelas il faut écrire & prononcer, Faire alte ; mais l'Academie est d'un autre sentiment, & veut qu'on dise, Faire halte. Richelet admet l'un & l'autre. Halte, se dit par extension, quand on s'arrête en faisant quelque chemin. […] (FURETIÈRE, 1701, t. II)[8]

Claude Favre, seigneur de Vaugelas (1585-1650) was a founder member of the *Académie Française* and was the prime mover on work on the dictionary up to his death in 1650. As such, Furetière both knew and cited him.

The lexicographical hourglass thus opens with the sources used by Basnage, which are cited as references as above, or directly coped in so as increase coverage in specialised areas.

Fig. 2: The lexicographical sources



To the left, we have the lexicographical sources, some of which, such as Isidore, are indirectly quoted. Amongst the dictionaries, we find that of the Academy (ACADÉMIE FRANÇAISE, 1694) and that of Thomas Corneille (CORNEILLE, 1694), published hurriedly to counter that of Furetière. Whilst the Academy dictionary retained only 'polite' language, the accompanying Corneille contained terms from the arts and sciences so as to offset the encyclopaedic side of Furetière's DU. Corneille made heavy use for Richelet's 1680 dictionary, a work that got past the Academy's monopoly on dictionary writing by having been published in Geneva. A variety of other works were also consulted.

[8] In English: "HALTE. Feminine noun. Aspired h. War term. […] Some say this is a German word, since they use halten with the same meaning. Vaugelas says it has to be written and pronounced, Faire alte; but the Academy considers otherwise, and recommends, Faire halte. Richelet uses both. Halte means in general a stop one makes when travelling. […]"

To the right, we have his *etymological* sources with Menage and Du Cange supplying the majority of his word histories. We then have a series of terminological inputs including a dictionary of maritime terms (DESROCHES, 1697), a book specialised in architecture (SAVOT, 1624), another one on blacksmith craft (SOLLEYSEL, 1654), and so on. A third section consists of philological discussions, inspired by writers as the couple Anne (1645-1720) and André Dacier (1651-1722).

Whilst these sources were often cited in the dictionary text, definitions from Desroches were often simply copied in. (FENNIS, 1988) mapped the sources of maritime terminology in the DU, noting the heavy use of Desroches. This is currently being reanalysed using the digital versions so as to compare the 1690 DU, where some definitions are close to that of Desroches, to the 1701 edition where Basnage sometimes copied Furetière verbatim, sometimes copied in Desroches verbatim, and sometimes, as with the entry for AMURES, rewrote the entry both for the spelling and using Desroches to add details. The italics show what has been retained from Furetière.

> Terme de Marine. *Ce sont des trous pratiquez dans le platbord d'un vaisseau, & dans la gorgere de l'éperon, pour y arrêter les cordages qui servent à bander les voiles.* Les amures des voiles d'étay, sont de simples cordes. *Les amures de la grande voile s'appellent dogues d'amures. L'amure d'une voile est son étoit, ou la manoeuvre qui sert à l'amurer. L'amure d'artimon, est un palanquin, ou quelquefois une corde simple.* L'amure à basbord ou à tribord, c'est à droit ou à gauche. (FURETIÈRE, 1701, t. I)[9].

Basnage was forced to work quickly and so the technique of copy-paste and comment was one he practiced often. In this case, he made use of an existing dictionary, but other sources are legion and the building of a virtual library for Basnage is a possibility, and one we are exploring.

In the prefaces, from that of Bayle onwards (WILLIAMS *et al*., 2020) mention is made of having made use of the finest authors. The *Vocabulario* of the *Accademia della Crusca* had done this, and so had Richelet, but the *Académie Française* had decided against, what is different in the usage of Furetière is that these are not simply to provide prestige to the entry, but are used to illustrate and bring in encyclopaedic information. Once CamemBERT is operational on the 1690 and 1701 editions we shall be able to compare sources, but it is quite clear that whilst Furetière relied heavily on his ex-circle in the *Académie Française*, Basnage was spreading his net far wider and tapping into the knowledge base of the protestant diaspora. This coupled with his European wide correspondence, his editorship of the journal *Histoire des ouvrages savants* and him being, according to Leibniz, a "'secrétaire d'état' de la République des Lettres" (GRAVELEAU, 2018, p. 295), Basnage drew his information widely. To date, we

[9] In English: Maritime term. These are holes made on the side of the ship, and to the front part, so as to place the ropes used to hoist the sails. The amures of shear sail are simple ropes. The amures of the great sail are called dogues d'amures. The amure of a sail is the manoeuvre that allows to hoist it. The amure of the mizzen is a pulley, or sometimes a simple rope. The larboard amure or starboard is right or left.

found some 240 persons cited, and these represent the best of seventeenth century science, each with several highly relevant works to their name.

There is thus a vast knowledge base at work with Basnage either adding in information or calling upon specialists who entirely rewrote the entries from Furetière and added in others. As we shall see in the following section, the botanical hourglass, major rewriting was essentially the work of other contributors, and notably Regis.

Input-Output: The botanical hourglass

Whilst the move to encyclopaedic knowledge is a long-term influence, a more immediate output comes through the mediation of information through detailed entries. As mentioned earlier, Basnage did what Furetière had not succeeded in doing and produced a truly encyclopaedic dictionary by calling on expert input on scientific matters and on matters of language. We shall concentrate on the input from the medical practitioner Regis, and in particularly what he brought in through studies of flora and fauna.

A good point of departure for the exploration of entries concerning the natural sciences can be seen in the first part of the entry for a Brazilian animal, the Coati.

> COATI. s. m. C'est un animal du Bresil diversement decrit par les Naturalistes, qui a un museau long d'un pied, rond comme un bâton, à peu-près comme la trompe d'un élephant, comme disent De Léri & Marcgravius. Cependant il n'en a rien que la mobilité ; car il ressemble davantage à un grouin de pourceau. De Laet en fait deux especes : l'un qui a le poil roux par tout le corps, & est appellé simplement coati ; & est la femelle : l'autre qui n'a que le ventre & la gorge de cette couleur, qu'on appelle coati mondi. On en a dissequé un de cette espece à l'Academie des Sciences, qui avoit six pouces depuis le bout du museau jusqu'à l'occiput, qui en avoit 16. jusqu'à la queuë, laquelle en avoit 13.de long.Il étoit haut de dix pouces. (FURETIÈRE, 1701, t. I)

The first issue here is why he added so many entries relating to Brazil. Brazil, Bresil in the dictionary, gets mentioned 98 times[10]. This is considerable for a distant country and for flora and fauna that are generally designated by the local native population, Tupi, name, which can hardly be considered as in general or technical usage in France. The answer partially lies in the author cited and in Regis, the likely compiler of these entries.

As we have shown elsewhere (GRAVELEAU *et al*., 2021), there are two main reasons for the interest in Brazil, and both lie in the existence of a short-lived, 16th century French colony on Brazilian territory called *La France antarctique* and the existence of a similar short-lived Dutch colony in the 17th century. Both Basnage and Regis being French would certainly have known of the writings of the French priest Thevet and the protestant minister de Léry,

[10] The information was obtained using SketchEngine on the partially cleaned OCR, so that the figures are not definitive. For term extraction, we use BaseX on the preliminary XML-TEI version so the same caveat applies.

both of whom had visited the colony and described life there. The fact that there was a controversy over different appreciations of the local tribes with Thevet seeing them as savages and cannibals, and de Léry describing them as noble savages albeit mistaken in religion is, also a factor, especially with Basnage and Regis also being protestant. In the above example, de Léry's work *Histoire d'un voyage fait en la terre du Brésil* (LÉRY, 1578) is cited a source. The other source mentioned, Marcgravius, concerns the publication following a voyage of exploration to the Dutch colony undertaken by Willem Piso, who brought with him the naturalist Georg Margraf Marcgravius, who, in collaboration with Piso and De Laet, published a major work on the natural life of Brazil, *Historia Naturalis Brasiliae… in qua non tantum plantae et animalia, sed et indigenarum morbi, ingenia et mores describuntur et iconibus supra quingentas illustrantur* (MARCGRAVIUS; PISO, 1648). Our work suggests that the latter work, and numerous others formed the library of Regis, that it was he who chose the words to be entered and who compiled the entries (GRAVELEAU *et al*. 2021).

In the above-mentioned paper, we looked only at trees, which represent 31 entries. However, for these entries, we found some 12 different sources[11], all of whom have clearly been consulted in building the descriptions. Amongst the sources we find the trio Piso-Marcgravius-de Laet from which Regis translated sections from the Latin text. De Léry is another source, as was the French-speaking botanical school associated with, Gaspard Bauhin. An excellent example of compilation is the entry for JABANIPA. This is a fairly long entry for one of the larger trees found in Brazil. The entry itself has no source author mentioned so it is necessary to look at how the tree is described in source texts already located. Analysis shows that the use made of the juice of the fruit in body painting comes from (THEVET, 1558), but that the wording does not[12]. The actual description of the tree is drawn from (LAET, 1640) in French and from the Latin texts of (MARCGRAVIUS; PISO, 1648). In this way, what we have in the entry is condensed information from notable travel and naturalist sources so that the dictionary effectively mediates between learned sources and a knowledge base for the user.

Thus, whilst the lexicographical sources are situated at the top of the hourglass, the botanical hourglass shows input and output from sources to the immediate availability of data to a reader. Now we can now turn to how the knowledge spread.

---

[11] See (Graveleau, Williams, et Villalva 2021) for details, but the following list reproduced from a note in the paper shows the twelves sources in chronological order: 1) (Thevet 1558), 2) (Léry 1578), 3) (Clusius 1611) , 4) (Bauhin 1623), 5) (Parkinson 1640), 6) (Laet 1640), 7) (Marcgravius et Piso 1648) (the texts of Piso and Marcgravius were published together. Following the death of Marcgravius in 1644, Piso simply removed his name and claimed authorship of the whole. He did the same to de Laet who died in 1649), 8) (Rochefort 1658), 9) Jacques de Bondt ou Bontius, *medici civitatis Batavia Nova*… published in a work under the name of Piso in 1658, 10) (Tertre 1667) – 11) (Ray 1688) – 12) (La Loubère, de 1691). To these twelve, he also consulted the (*Dictionnaire de l'academie francoise* 1694) and the Dictionnaire des sciences et des arts of (Corneille 1694).

[12] "Le suc de ce fruit est blanc d'abord, & quand on s'en est frotté le corps, il noircit en peu de temps de telle sorte que les Sauvages s'en servent au lieu d'ancre, pour paroître plus terribles à leurs ennemis" (t. II). Translation: "the juice from this fruit is at first white, and when it is rubbed on the body, it blackens in a short while so that he savages use it instead of ink so as to appear more frightening to their enemies."

Hourglass output: the influence of the *Dictionnaire Universel*

This is a much more complex situation and one that is the subject of ongoing research. The various editions of the DU were involved in an interesting interplay with rival productions. Thus, the DU 1690 forced the Académie-Française into action so that they finally produced the dictionary that was to be their main task at their foundation 1635 some 60 years later in 1694. This latter publication resulted in the Dutch publisher Leers inviting the protestant exile Henri Basnage de Beauval to produce a new version in 1701. This version was heavily revised and greatly increased in length and coverage. This was in turn plagiarised by the Jesuits of Trévoux in 1704, resulting in a marginally updated version from Basnage in 1708. It is at this point that the Trévoux dictionary began to break away from simply copying and moved towards the more encyclopaedic editions that followed.

Picking up on Basnage's comment cited above as to the use of specialised contributors, the Jesuits claimed to be doing likewise, and to be the first so to do, although this seems to have simply been rhetoric (LECA-TSIOMIS, 1999). Nevertheless, the 1721 Trévoux meant that Brutel de la Rivière, the successor to Basnage who died in 1710, had to take up again the revision that he had started earlier so that a new, and final version, of the DU was published in 1725/27. From this point on, the Trévoux Jesuits has a new enemy to fight, the humanists Diderot and d'Alembert. This then is one major use of the dictionary, albeit not cited as such.

In a similar way, reeditions of earlier dictionaries can also have been updated using, but not acknowledging, the 1701 DU. A case in point is the entry for 'Celestin' where he adds a story to illustrate the usage "voilà un plaisant Celestin" (this is a funny Celestine) an expression that is generally attributed to Richelet for its first attribution, but which is not to be found in the 1680 Richelet, but only in the 1724 version. (GRAVELEAU, forthcoming) Thus, the indirect influence is clearly present.

The 1690 dictionary was a great success and was part of the library of the great Danish lexicographer Matthias Moth (1649-1719) who was engaged on his *Ordbog* (EEGHOLM-PEDERSEN, 2018). Moth apparently had no knowledge of the 1701 edition.

In terms of a direct heritage of the DU, we find the work of Rafael Bluteau (1638-1734) a French Jesuit born in London, but naturalized Portuguese. Bluteau's *Vocabulario* is the first major dictionary of Portuguese, and is also an encyclopaedic dictionary of the model of the DU. Bluteau has been inspired by the 1690 DU and it is this of which he made the most use when compiling his own dictionary (SILVESTRE, 2008). In his list of sources, in addition to DU 1690, Bluteau cites Richelet and the Academy and also a *Dictionnaire Universel*, 1709. Insofar as the second Basnage was published in 1708, this must be either a mistake, or a pirate edition.

Silvestre demonstrates the reliance of Bluteau on Furetière through the example of the entry for DIAPHRAGME*,* in which the wording in the 1690 DU (s. p.) is: "Terme de Medecine. Membrane ou muscle nerveux qui separe la poitrine d'avec le bas ventre, & qui est comme une espece de plancher qui est entre les parties vitales & les naturelles." This corresponds closely to that of Bluteau:

> DIAFRAGMA. (Termo Anatomico.) Derivase do Grego Diaphratein, que val o mesmo, que dividir huma cousa da outra, como frontal, ou muro divisorio. O diafragma, he hum paniculo, ou membrana musculosa, que atravessando o peito, divide, & separa os membros vitaes, a saber, o coraçaõ, & os bofes, dos membros naturaes, a saber, o baço, & intestinos. (FURETIÈRE, 1701, s. p.)

The entry for diaphragm found in the 1701 DU had been heavily revised, presumably by Regis, so it is quite clear that Bluteau had relied on the earlier edition. Insofar as Bluteau's work was the basis for future encyclopaedic dictionaries of Portuguese, we see an ongoing inheritance that predates the Basnage edition, and hence also the pirate Trévoux dictionary.

The move to the domination of the encyclopaedia is also interesting to follow. Harris' *Lexicon technicum* of 1708 (HARRIS, 1708) did draw on Furetière, and this in turn inspired Chambers. In his preface, Harris admitted having consulted Furetière, but claimed it had been of little use to him: "And therefore, this and Mr *Furetière's Dictionary* may be Books well done in their way, and are certainly very useful for those who would be acquainted with the *French* Tongue; yet I did not find much assistance from them, […]" (HARRIS, 1708, s. p.)

He details other works consulted, and notably the *Dictionnaire des arts et des sciences* (CORNEILLE, 1694) which he attributes to the Académie-Française although it was in fact a companion volume to their dictionary of the French language, and one designed to offset the encyclopaedic aspect of Furetière's work. For Harris, Corneille's work listed far too many non-scientific words and was primarily a tool "to improve and propagate the French language, than to inform and instruct the Humane mind in general". He also quotes more specialised works as the *Dictionnaire mathématique* of (OZANAM, 1691). From his comments, it seems clear that it was the 1690 version of the DU that he was using and not that of Basnage, he was thus also unaware of the Trévoux.

Whilst Harris had relatively limited aims, as Chambers notes in his preface (CHAMBERS, 2020, s. p.) he sought universality and thus indirectly refers to the model that had been created by Furetière where he claims that he will create a "Cyclopaedia, or, An universal dictionary of arts and sciences: containing the definitions of the terms, and accounts of the things signify'd thereby, in the several arts, both liberal and mechanical, […]".

On the other hand, whilst Chambers also lists sources, he mentions the Trevoux and not the Furetière: "What the French Academists, the Jesuits de Trevoux, Daviler, Chomel, Savary, Chauvin, Harris, Wolsius, and many more have done, has been subservient to my Purposes." The Academists presumably means the Corneille dictionary. Bocast (CHAMBERS, 2020, p. 5) believes that Chambers primarily used the 1721 Trévoux, although many of his definitions were verbatim from the 1701 DU, mediated by the 1704 Trévoux plagiarism.

In the Catholic world, unsurprisingly, it is the Trévoux version that spread the most, and thus the influence is assigned to the 1704 dictionary. Insofar as the latter was a simple copy, when we find mention in the *Diccionario Castellano con voces de ciencias y artes* (1786-1788) and the *Diccionario Italiano, e Portuguez* (1773-1774) of Joaquim José da Costa e Sá what

they are citing is the work of Basnage and his contributors[13]. This is a hidden hourglass effect, but an important one given the prevalence of the encyclopaedic dictionary in the 18th and 19th centuries.

The dominance of the encyclopaedic dictionary was obviously weakened by the rise of genuine encyclopaedias from that of Alembert and Diderot onwards. These more specialised works deal with ideas and things more than words, but the inheritance of the *Dictionnaire Universel* is clear from the *Avertissement* to the third volume (DIDEROT; D'ALEMBERT, 1751) where d'Alembert states in no uncertain terms that:

> Le Dictionnaire de Trévoux en particulier doit moins reprocher qu'aucun autre les *emprunts* à l'Encyclopédie; car ce Dictionnaire n'étoit dans son origine & n'est encore en grande partie, qu'une copie du Furetiere de Basnage, ainsi que ce dernier l'a fait voir & s'en est plaint dans son histoire des ouvrages des Savans (*n*).
>
> (The Trévoux dictionary in particular should not reproach more than any other to be borrowings of the Encyclopaedia as this dictionary was from its beginnings, and is largely still, just a copy of the Furetière of Basnage, as the last-mentioned saw and complained in his *Histoire des ouvrages des savans*). (DIDEROT; D'ALEMBERT, 1751, p. 3: viij)

As is pointed out, the fact that the *Dictionnaire de Trévoux* is more known explains why later editions of the 'Furetière' are less known. The Jesuits had done the utmost to hide their intellectual theft, but when accused of having had recourse to the 1748 Trevoux in compiling the encyclopaedia (LECA-TSIOMIS, 1999), the compilers are keen to re-establish the truth in writing the entry for *Buses (hydraulique)* they note that:

> Nous avons averti que le Dictionnaire de Trévoux est en grande partie copié du Furetiere de Basnage. Ainsi quand nous citerons dans la suite le Dictionnaire de Trévoux, c'est seulement parce que le nom de celui - ci est plus connu, & sans prétendre faire tort à l'autre qui a été son modele. Plusieurs des articles de l'Encyclopédie qu'on a prétendu être imités ou copiés du Trévoux, sont eux - mêmes imites ou copiés de Basnage […]
>
> (We have warned that the Trévoux dictionary is to a great extent copied from the Furetière of Basnage. Thus, when in the following we cite the Trévoux dictionary,

[13] A workshop on "Lexicographical networks in the 17th and 18th centuries" was held in Paris on the 31st January 2020. Speakers included Monica Lupetti (University of Pisa) on *Il* Diccionario Italiano, e Portuguez *(1773-1774) di Joaquim José da Costa e* and Elena Carpi, (Università di Pisa), Francisco M. Carriscondo Esquivel (uma - iemyrhd - anle), *Anglicismes introduits par le français dans le* Diccionario Castellano con voces de ciencias y artes *(1786-1788)*. Joao Silvestre, (Kings College London) presented Blueau's work as *Furetière as an omitted source in Portuguese lexicography* Kira Kovalenko, & Georgiy A. Molkov (Russian Academy of Sciences) described *Russian Multilingual Dictionaries of the 18th C.: Relations with French Lexicography*, the influence of Furetière and Basnage in Russia. An unfinished encyclopaedia, and hence another rival to the Trévoux was presented by Linn Holmberg, (Stockholm University), *The Maurists' unfinished dictionary.* The lexicographical background to Diderot and Alembert was provided by Marie Leca-Tsiomis, (Université Paris-Ouest Nanterre), *La genèse lexicographique de l'*Encyclopédie. Alexander Bocast was unable to attend, but his two works on the influence on Chambers are cited above. A book is to be published with these papers.

> it is only because the name of this one is more widely known and not to bring damage to that which was its model. Several entries of the Encyclopaedia have been claimed to have been imitated or copied from the Trévoux, which were themselves imitated or copied from Basnage […]) (LECA-TSIOMIS, 1999, p 3: xvj)

Conclusion

Encyclopaedic dictionaries were a relatively short lived, but highly influential phenomenon. The period starts with the publication of Furetère's *Dictionnaire universel* (FURETIÈRE, 1690) with a much enlarged and revised edition by (FURETIÈRE, 1701) and gradually dies out in the early nineteenth century after the rise of domination of the encyclopaedia in the eighteenth century. From the mid-nineteenth century onwards, under the influence of the *Worterbuch* by the brothers Grimm, dictionaries turned to questions of philology and more concise definitions leaving the large knowledge bases to the encyclopaedia. Nevertheless, their influence was considerable and the simile of the hourglass illustrates how Basnage revolutionised the genre by calling upon specialist input that would mediate knowledge from a large number of learned sources and make it available in a condensed form to readers of their time. Also, the nature of the sources gives a deep insight into the state of the art in the sciences of the time.

References

ACADÉMIE-FRANÇAISE. *Dictionnaire de l'Academie Françoise*. Paris: J.-B. Coignard, 1694.

BAUHIN, G. *Pinax theatri botanici Caspari Bauhini,... sive Index Theophrasti Dioscoridis Plinii et botanicorum qui a seculo scripserunt opera*. Basel : L. Regis (Basileae Helvet), 1623.

BAYLE, P. Préface. In : FURETIÈRE, A. *Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes et les termes des sciences et des arts*. La Haye & Rotterdam: Arnout & Reynier Leers, 1690, s. p.

CHAMBERS, E. *A Circle of Knowledge for* Definition *in Chamber's* Cyclopaedia. Edited by Alexander BOCAST. Anacortes: Berkeley Bridge Press, 2020.

CORNEILLE, T. *Dictionnaire des Arts et des Sciences*. Paris: Veuve J.-B. Coignard et J.-B. Coignard, 1694.

DESROCHES, N. *Dictionnaire des termes propres de la marine*. Paris: Aimable Auroy, 1687.

DIDEROT, D.; D'ALEMBERT, J. (Eds.). Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers. Tome premier. Paris: Briasson, David l'aîné, Le Breton, Durand, 1751.

EEGHOLM-PEDERSEN, S. *Mothstudier: Kildegrundlaget for den første store danske ordbog*. Denmark: Universitets-Jubilæets danske Samfund, 2018.

FENNIS, J. Les sources du vocabulaire maritime dans le Furetière de 1701. *Travaux de linguistique et de philologie*, n. XXVI, v. 1, p. 75-94, 1988.

FURETIÈRE, A. *Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes et les termes des sciences et des arts*. La Haye & Rotterdam: Arnout & Reynier Leers, 1690.

FURETIÈRE, A. *Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes et les termes des sciences et des arts*. 2e edition, revue, corrigée et augmentee par M. Basnage de Beauval. La Haye & Rotterdam: Arnout & Reynier Leers, 1701.

GRAVELEAU, S. *'Les hérésies sont d'utiles ennemies'. Itinéraire d'Henri Basnage de Beauval (1656-1710), avocat de la République des Lettres et penseur de la tolérance civile*. Doctoral dissertation, Faculté d'Histoire, Université d'Angers, Angers (France), 2018.

GRAVELEAU, S. La Normandie et les Normands face à l'étranger dans le Dictionnaire universel d'Henri Basnage de Beauval. Le regard d'un enfant du pays porté depuis l'exil. *Annales de Normandie*. Forthcoming.

GRAVELEAU, S.; WILLIAMS, G. C.; VILLALVA, A. Les arbres du Brésil dans deux dictionnaires historiques: le Dictionnaire universel de Basnage et le Vocabulario de Bluteau. *TradTerm*. Forthcoming.

GRAVELEAU, S.; STINCONE, C.; WILLIAMS, G.; GALLERON, I. Linking Authors and Works in the Dictionnaire Universel by Basnage de Beauval (1701). *Linked Pasts 6*, London, Institute of Classical Studies - Schoo of Advanced Study University of London, 2020. Available at: https://ics.sas.ac.uk/events/linked-pasts-6/poster-session Last accessed: 25 jul. 2021.

HANKS, P. Typicality and meaning potentials. In: *ZüriLEX '86 Proceedings*. Zurich: Francke Verlag, 1986, p. 37-47.

HARRIS, J. *Lexicon technicum, or, An universal English dictionary of arts and sciences: explaining not only the terms of art, but the arts themselves.* London: Printed for Dan. Brown, Tim. Goodwin, John Walthoe, Tho. Newborough, John Nicholson, Dan. Midwinter, and Francis Coggan, 1708.

KHEMAKHEM, M.; FOPPIANO, L.; ROMARY, L. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *eLex 2017 Proceedings,* Leiden*:* U. of Leiden, 2017. Available at: https://hal.archives-ouvertes.fr/hal-01508868v1 Last accessed: 03 mar. 2021.

KHEMAKHEM, M.; GALLERON, I.; WILLIAMS, G. C.; ROMARY, L.; ORTIZ SUAREZ, P. J. How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures. *Proceedings of the 19th annual Conference and Members' Meeting of the Text Encoding Initiative Consortium (TEI) -What is text, really? TEI and beyond*, Graz: U. of Graz, 2019. Available at: https://hal.archives-ouvertes.fr/hal-02263276 Last accessed: 03 mar. 2021.

LA LOUBÈRE, S. de. *Du Royaume de Siam.* Paris: J.-B. Coignard, 1691.

LAET, J. de. *L'Histoire du Nouveau Monde ou Description des Indes occidentales*. Anvers: Bonaventure & Abraham Elsevier, 1640.

LECA-TSIOMIS, M. *Écrire l'Encyclopédie. Diderot: de l'usage des dictionnaires à la grammaire philosophique*. Oxford: Voltaire Foundation, 1999.

LÉRY, J. de. *Histoire d'un voyage fait en la terre du Brésil*. La Rochelle: Antoine Chupin, 1578.

MARCGRAVIUS, G.; PISO, W. *Historia Naturalis Brasiliae… in qua non tantum plantae et animalia, sed et indigenarum morbi, ingenia et mores describuntur et iconibus supra quingentas illustrantur*. Amsterdam: Elsevier, 1648.

ORTIZ SUARES, P. J.; DUPONT, Y.; MULLER, B.; ROMARY, L.; SAGOT, B. Establishing a New State-of-the-Art for French Named Entity Recognition. *Proceedings of LREC - 12th*

Language Resources and Evaluation Conference, Marseille, U. of Aix-Marseille, 2020. Available at : https://hal.inria.fr/hal-02617950 Last accessed: 03 mar. 2021.

OZANAM, J. *Dictionaire Mathematique ou Idée generale des Mathematiques*. Paris: Estienne Michalet, 1691.

PARKINSON, J. *Theatrum Botanicum = the Theater of Plants* : Or, An Herball of a Large Extent. London: Tho. Cotes, 1640.

REY, A. *Antoine Furetière : Un précurseur des Lumières sous Louis XIV*. Paris: Fayard, 2006.

ROCHEFORT, Ch. De. *Histoire naturelle et morale des Isles Antilles de l'Amerique.* Rotterdam: A. Leers, 1658.

ROY-GARIBAL, M. *Le Parnasse et le Palais. L'ouevre de Furetière et la génèse du premier dictionnaire encycloédique en language française (1649-1690)*. Paris: Honoré Champion, 2006.

RUNDELL, M. Defining elegance. In SCHRYVER, G.-M. de (Ed.) *A Way with Words*. Kampala, Uganda: Menha Publishers, 2010, p. 349-375.

SAVOT, L. *L'Architecture françoise des bastimens particuliers*. Paris: F. Clouzier, 1624.

SOLLEYSEL, J. de. *Le Parfait mareschal qui enseigne à connoistre la beauté, la bonté et les deffauts des chevaux, la manière de les conserver dans les fatigues des voyages*. Paris: G. Clousier, 1654.

SILVESTRE, J. P. *Bluteau e as origens da lexicografia moderna*. Lisboa: Imprensa Nacional – Casa da Moeda, 2008.

TERTRE, J.-B. du. *Histoire generale des Antilles habitées par les François*. Paris: Thomas Jolly, 1668.

THEVET, A. *Les Singularitez de la France antarctique, autrement nommée Amérique, et de plusieurs terres et isles découvertes de nostre tems*. Paris: Maurice de la Porte, 1558.

VORSTIUS, E. *Caroli Clusii Atrebatis Curae Posteriores…*, Leiden : in officina Plantiniana Raphelengii, 1611.

WILLIAMS, G. C.; GALLERON, I. Digitizing the second edition of Furetière's Dictionnaire Universel: challenges of representing complex historical dictionary data using the TEI. In: *Proceedings of the XVIIth EURALEX International Congress*. Tbilissi: Ivane Javakhishvili Tbilisi University Press, 2016, p. 647-652.

WILLIAMS, G. C.; GALLERON, I.; STINCONE, C. Announcing the Dictionary: Front Matter in the Three Editions of Furetière's Dictionnaire universel. In: *Proceedings of the XIXth EURALEX Congress*. Alexandropoulis: Democritus University of Thrace, 2020, p. 393-402.

WIONNET, C.; TUTIN, A. *Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704) : Approche théorique et pratique*. Paris: Honoré Champion, 2001.

Submitted: 05/17/2021

Accepted: 06/02/2021

Linha D'Água: São Paulo, v. 34, n. 02, p. 26-46, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p26-46

Artigo / Article

Terminologia e organização do conhecimento: linguagens, vocabulários e sistemas

Terminology and Knowledge Organization: Languages, Vocabularies and Systems

Bruno Almeida*

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

Resumo

Este artigo parte do pressuposto de que as ferramentas utilizadas para a organização do conhecimento (tesauros, esquemas de classificação, etc.) podem ser entendidas como recursos terminológicos. Abordamos as relações entre terminologia e organização do conhecimento, assumindo um ponto de vista baseado na bidimensionalidade (linguística e conceptual) da terminologia enquanto disciplina. Em seguida, propomos uma análise dos conceitos subjacentes às designações "linguagem documental", "vocabulário controlado" e "sistema de organização do conhecimento" nos textos de especialidade. Terminamos com a descrição do SKOS (*Simple Knowledge Organization System*), um modelo para a representação de sistemas de organização do conhecimento na *web* semântica, o qual é avaliado em termos da sua capacidade de modelizar recursos terminológicos de acordo com a abordagem bidimensional à terminologia e os principais elementos da norma ISO 1087.

Palavras-chave: Sistemas de organização de conhecimento; Vocabulários controlados; Linguagens documentais; Recursos terminológicos; SKOS.

Abstract

This article assumes that the tools used for knowledge organization purposes (e.g., thesauri, classification schemes…) may be understood as terminological resources. We investigate the relationship between terminology and knowledge organization from the double-dimension perspective (both linguistic and conceptual) of terminology as a field of study.

* Centro de Linguística da Universidade NOVA de Lisboa, Portugal.

Then, we propose an analysis of the concepts underlying designations such as 'documentary language', 'controlled vocabulary' and 'knowledge organization system' in specialized texts. We conclude with an overview of SKOS (Simple Knowledge Organization System), a model to represent knowledge organization systems in the semantic web, which is evaluated in terms of its ability to model terminological resources according to the double-dimension approach and the main elements of the ISO 1087 standard.

Keywords: *Knowledge Organization Systems; Controlled Vocabularies; Documentary Languages; Terminological Resources; SKOS.*

Introdução

A linguística e a ciência da informação são duas áreas de conhecimento distintas, ainda que estejam relacionadas a diversos níveis, sobretudo no atual panorama académico marcado pela trans- e interdisciplinaridade. Uma das relações de maior relevo entre estas áreas pode ser estabelecida através da terminologia, entendida aqui como área de estudos de natureza bidimensional, linguística e conceptual, assumindo um papel de *disciplina de interfaces* entre as ciências da linguagem e as diversas áreas de conhecimento, em particular no que diz respeito ao léxico e às línguas de especialidade (COSTA; SILVA; BATISTA, 2020).

Deste ponto de vista, a terminologia relaciona-se com a *organização do conhecimento* (OC), um subdomínio da ciência da informação. A organização do conhecimento, ou para usar uma designação mais recente, *sistemas de organização do conhecimento* (SOC) (HJØRLAND, 2008; HODGE, 2000; ZENG, 2008), consiste na prática e estudo de atividades como a indexação por assuntos e classificação documental, ou a construção e aplicação das chamadas *linguagens documentais* e *vocabulários controlados*. Estes recursos incluem as classificações e listas de cabeçalhos de assuntos, muito utilizadas em bibliotecas, arquivos e outros serviços de informação, embora atualmente incluam tipos de recursos surgidos em ambiente digital, como as ontologias da representação do conhecimento e da *web* semântica (GUARINO; OBERLE; STAAB, 2009). A adoção das tecnologias de informação e o desenvolvimento da *web* levaram a transformações significativas nas tradicionais linguagens documentais, deixando estas de estar limitadas ao suporte em papel. A transição para o meio digital tornou viável, entre outros aspetos, o desenvolvimento de recursos de grandes dimensões e complexidade, abrangendo dezenas de milhares de conceitos com designações em diversas línguas, assim como a sua aplicação no enriquecimento e descoberta de dados em bibliotecas digitais e portais semânticos.

Neste artigo, assumimos que os SOC são recursos terminológicos e, como tal, podem ser abordados tendo em conta a bidimensionalidade da terminologia enquanto área de estudos das ciências da linguagem. Propomos, em seguida, uma análise dos conceitos acima referidos, desde linguagens documentais a vocabulários controlados e SOC, terminando com uma análise do SKOS (*Simple Knowledge Organization System*), um modelo para representação de SOC na *web*.

1 Terminologia e organização do conhecimento

1.1 A terminologia enquanto área de conhecimento

O lexema *terminologia*, assim como os seus equivalentes noutras línguas, apresenta diversos significados, que abrangem o vocabulário utilizado em áreas especializadas, assim como a disciplina dedicada à recolha e estudo desse vocabulário. Neste sentido, Sager (1990) classificou o equivalente inglês, *terminology*, como termo polissémico impróprio (*polysemous misnomer*), uma vez que nenhum dos seus significados corresponde precisamente ao seu significado etimológico:

> By its etymology 'terminology' would mean 'the science/study/knowledge of terms' which would make it parallel to lexicology, the science/study/knowledge of the lexicon or lexical items; this interpretation is, however, rejected by most terminologists. (SAGER, 1990, p. 3)

De facto, apesar da pluralidade de perspetivas sobre a terminologia enquanto área de conhecimento (CABRÉ, 2003; COSTA, 2006; FABER, 2012; FELBER, 1984; ISO 1087, 2019; ROCHE, 2007; TEMMERMAN, 2000; WÜSTER, 1979), verificamos que as diversas abordagens teóricas não identificam a terminologia com o estudo do léxico das línguas. Pelo contrário, estas abordagens reconhecem os seguintes aspetos:

- A terminologia envolve contributos teóricos e metodológicos de diversos domínios, tais como a linguística, a filosofia, a psicologia, a informática, entre outros.
- A terminologia é caracterizada pela complexidade das unidades terminológicas, as quais podem ser abordadas segundo múltiplos pontos de vista e a partir de diversas disciplinas.

Apesar dos diversos pontos em comum, não podemos deixar de salientar as diferenças mais significativas entre as diversas abordagens à terminologia:

- O posicionamento da terminologia, quer como ciência mais ou menos autónoma (ISO 1087, 2019; ROCHE, 2007; WÜSTER, 1979), quer como área de estudos interdisciplinar (CABRÉ, 2009; FABER, 2012; SAGER, 1990; TEMMERMAN, 2000).
- O foco na normalização de termos e conceitos (FELBER, 1984; WÜSTER, 1979) ou na descrição do uso dos termos em textos de especialidade (CABRÉ, 2009; FABER, 2012; SAGER, 1990; TEMMERMAN, 2000).
- Ênfase nos termos (CABRÉ, 2009; SAGER, 1990; TEMMERMAN, 2000) ou nos conceitos (ISO 1087, 2019; ROCHE, 2007; WÜSTER, 1979), enquanto ponto de partida do trabalho terminológico e/ou elemento teórico central.

- Conceção semiótica dos termos enquanto designações atribuídas convencionalmente a conceitos previamente definidos (ISO 1087, 2019; ROCHE, 2007; WÜSTER, 1979), ou conceção linguístico-discursiva dos termos, determinada pelo seu uso em textos de especialidade (CABRÉ, 2009; FABER, 2012; SAGER, 1990; TEMMERMAN, 2000).

Tendo em conta a natureza interdisciplinar deste domínio, devemos também salientar a diversidade de métodos utilizados na investigação e no trabalho terminológicos, que vão desde a linguística de *corpus* (BOWKER; PEARSON, 2002; MELBY, 2012) ao trabalho colaborativo e validação de dados terminológicos envolvendo especialistas de domínio (SILVA, 2014).

A diversidade de abordagens teóricas em terminologia manifesta-se de forma mais acentuada no que diz respeito aos aspetos que incluímos na dimensão conceptual, tais como conceitos, sistemas conceptuais e modelos para representação de conhecimento. A primeira abordagem teórica em terminologia, a teoria geral da terminologia (FELBER, 1984; WÜSTER, 1979), colocou a ênfase na dimensão conceptual como traço distintivo da terminologia em relação à linguística estruturalista. Wüster (1979) e Felber (1984) salientaram os contributos de áreas como a lógica, a ontologia e a ciência da informação na construção de sistemas conceptuais para o trabalho terminológico. Esta abordagem tem como descendentes diretos, atualmente, o Comité Técnico 37 da Organização Internacional de Normalização (ISO), dedicado à redação de normas internacionais em terminologia (ISO 704, 2009, p. 70; ISO 1087, 2019), e a ontoterminologia, que coloca a ênfase no desenvolvimento de recursos terminológicos baseados em ontologias para a representação do conhecimento (ROCHE, 2007).

Desde a década de 1990, várias abordagens teóricas procuraram posicionar a terminologia como área de estudos da linguística, embora reconhecendo a sua natureza interdisciplinar. É este o caso da teoria comunicativa da terminologia (CABRÉ, 2000, 2009), que elege o termo como elemento teórico central, concebido como unidade linguística, cognitiva e sociocomunicativa. O termo é abordado como uma unidade lexical que adquire um valor especializado, ou *terminológico*, no léxico dos especialistas de domínio. O conceito, elemento distintivo da terminologia wüsteriana, passa a ser entendido como um aspeto da unidade terminológica, podendo ser abordado como objeto de estudo da semântica lexical.

A teoria comunicativa da terminologia abriu caminho para abordagens baseadas em enquadramentos teóricos da semântica cognitiva, nomeadamente a terminologia sociocognitiva (TEMMERMAN, 2000) e a terminologia de marcos semânticos (*frame-based terminology*) (FABER, 2012). No caso da terminologia sociocognitiva, é dado destaque à metáfora conceptual e à semântica prototípica (LAKOFF, 1987; LAKOFF; JOHNSON, 1980) na descrição de unidades lexicais especializadas, enquanto que na terminologia de marcos semânticos o significado dos termos é descrito através de padrões de conhecimento baseados em marcos semânticos (FILLMORE, 1985).

1.1.1 A abordagem bidimensional à terminologia

A abordagem bidimensional tem procurado afirmar-se nas últimas décadas como enquadramento teórico e metodológico baseado na distinção entre as dimensões linguística e conceptual da terminologia (COSTA, 2006, 2013; COSTA; SILVA; BATISTA, 2020; SANTOS; COSTA, 2015). Assim, partimos do pressuposto de que a terminologia é uma área de estudos interdisciplinar e que, como tal, nasce da confluência entre diversas áreas de conhecimento, incluindo vários subdomínios da linguística, filosofia e informática.

A distinção entre as dimensões linguística e conceptual no plano teórico facilita a conciliação de métodos semasiológicos, centrados na constituição e análise de *corpora* de textos de especialidade, e métodos onomasiológicos, centrados na elaboração de sistemas conceptuais. Enquanto a extração e o estudo de termos, colocações e marcadores linguísticos relevam da dimensão linguística, a modelização do conhecimento releva da dimensão conceptual, que é – por definição – extralinguística. O texto de especialidade é o elemento central do trabalho terminológico, já que ambas as metodologias requerem a certo ponto o recurso aos textos produzidos por especialistas e disseminados através das suas comunidades de prática, seja no meio académico, técnico ou artístico.

Em suma, a abordagem bidimensional baseia-se nos seguintes pressupostos:

- Reconhecimento da *natureza interdisciplinar* da terminologia enquanto área de estudos.
- Distinção entre uma *dimensão linguística* e uma *dimensão conceptual* na terminologia e no trabalho terminológico.
- Confluência de métodos *semasiológicos* e *onomasiológicos*.
- Papel central do texto de especialidade.

No contexto da investigação em semântica, o postulado de um nível de análise linguístico e um nível de análise conceptual não é novo. Encontramos eco deste postulado, por exemplo, em Wierzbicka (1996), quando defende a distinção entre o significado lexical e o conhecimento científico. Para esta autora, a distinção justifica-se pelo facto de as línguas e o conhecimento constituírem dois planos de análise distintos:

> Science is, or tries to be, universal and to reflect the knowledge accumulated by mankind as a whole (and, more specifically, by the professional experts in different fields of knowledge); languages are not universal, and each of them reflects the experience of a particular part of mankind, united by a common culture and a common existential framework (WIERZBICKA, 1996, p. 338)

Seguindo o mesmo postulado, não identificamos a análise do significado de unidades lexicais especializadas com a organização do conhecimento, uma vez que esta última não releva de nenhum sistema linguístico. Este postulado facilita o trabalho de natureza trans- e interdisciplinar em terminologia, incluindo a constituição de *corpora* de especialidade e a

extração semiautomática de termos (ALMEIDA; COSTA; ROCHE, 2019), a relação entre a terminologia e as ontologias da representação do conhecimento (ALMEIDA; COSTA, 2021) e a construção de vocabulários controlados e sua publicação na *web* como dados abertos e ligados (ALMEIDA; FREIRE; MONTEIRO, 2021).

Neste capítulo, descrevemos sumariamente a terminologia enquanto área de conhecimento e posicionámo-nos no contexto das diversas abordagens teóricas e metodológicas que têm vindo a ser propostas nas últimas décadas. Apresentaremos, em seguida, a OC como subdomínio da ciência da informação, tendo em conta ainda a sua relação com a terminologia.

1.2 Organização do conhecimento (OC)

De acordo com Hjørland (2008), a OC pode ser definida através dos seguintes objetos de estudo:

- *Processos de organização do conhecimento*, tais como a indexação de documentos por assuntos, isto é, a atribuição de descritores, ou termos de indexação, aos documentos e a classificação de documentos por temática ou área de conhecimento. Estes processos são tradicionalmente levados a cabo em bibliotecas, arquivos, museus e outras instituições de memória.
- *Sistemas de organização do conhecimento* utilizados nos processos acima referidos para organizar os documentos, as suas representações, as obras que os abrangem e os próprios conceitos das diversas áreas de conhecimento.

Este autor salienta, todavia, que não é possível fazer uma reflexão teórica rigorosa neste domínio isoladamente. Pelo contrário, a OC deverá ser abordada num sentido mais amplo, abrangendo o estudo da produção e disseminação do conhecimento nas sociedades. Hjørland (2008) nota que diversos domínios contribuem para esta abordagem interdisciplinar à OC, como por exemplo, a linguística, a psicologia e a sociologia. Apesar das diferentes metodologias e perspetivas disciplinares que separam estas áreas, Hjørland (2008) enfatiza a existência de um plano epistemológico comum, por exemplo, ao nível das teorias do conceito, as quais podem sustentar a investigação nas diversas disciplinas.

1.2.1 A teoria do conceito em OC

É na teoria do conceito que encontramos a relação mais direta entre terminologia e OC, sobretudo através do trabalho de Dahlberg (1978, 1992, 2009). Esta autora propôs a chamada *teoria analítica do conceito* (*referent-oriented, analytical concept theory*), a qual ainda hoje transparece nas normas ISO sobre o vocabulário, princípios e métodos em terminologia (ISO 704, 2009; ISO 1087, 2019). Como refere Campos (2001), a teoria analítica do conceito foi inicialmente proposta como base para a terminologia das ciências sociais, em reação ao peso

excessivo da engenharia e da normalização no trabalho de Wüster (1979). Em seguida, a teoria proposta por Dahlberg (1992) assumiu um papel central nos estudos sobre classificação, estendendo-se mais tarde à OC.

Segundo Dahlberg (1992), o conhecimento é um estado de consciência em relação ao mundo, o qual se manifesta através da linguagem, mais especificamente, através de proposições verdadeiras sobre os referentes no mundo. A natureza analítica desta teoria deve-se à noção de que o conhecimento pode ser decomposto em *conceitos*, os quais correspondem a combinações únicas de *características* ou elementos conceptuais:

> One cannot predicate a true fact without expressing one's knowledge of something. Thus every predication yields a knowledge element and the necessary sum total of predications can be synthesized into the corresponding knowledge unit. Concepts are thus the units of our knowledge, and this is an essential finding on which we can base all our activities in the area of knowledge organization and terminology. (DAHLBERG, 1992, p. 66)

Tanto a terminologia como a OC preconizam a criação de sistemas conceptuais de forma a tornar explícito o conhecimento do domínio de análise e a apoiar a redação de definições. Os sistemas conceptuais consistem em diversos tipos de relações entre conceitos, hierárquicas e não hierárquicas, tais como genéricas, partitivas e diversos tipos de relações associativas (NUOPPONEN, 2014).

Por outro lado, os termos correspondem a *designações* ou *denominações* de conceitos. Como tais, Dahlberg (1992) considera que os termos devem refletir as *características essenciais* dos respetivos conceitos através dos seus constituintes morfológicos ou lexicais. Por outro lado, os termos devem facilitar a comunicação especializada, permitindo formar paradigmas derivacionais. A *internacionalização* dos termos é também encorajada, por exemplo, através de empréstimos linguísticos de termos utilizados internacionalmente.

Embora a teoria analítica do conceito seja útil para a elaboração de recursos terminológicos, devemos salientar a existência de outras abordagens ao conceito na investigação em OC. Neste particular, Hjørland (2009) identificou quatro pontos de vista epistemológicos:

- *Empirismo*. Os conceitos devem ser definidos através da aglomeração de objetos com propriedades em comum.
- *Racionalismo*. Os conceitos devem ser definidos através da combinação de *primitivos semânticos*, tais como facetas ou características conceptuais indecomponíveis.
- *Historicismo*. Os conceitos devem ser definidos pela sua genealogia em relação às teorias e discursos que os sustentam.
- *Pragmatismo*. Os conceitos devem ser definidos pela sua utilidade num sistema e fixados através de signos.

Como salienta Hjørland (2009), a investigação em sistemas de organização do conhecimento (SOC) exige o reconhecimento dos pontos de vista epistemológicos que subjazem a estes sistemas, uma vez que o conhecimento é sempre baseado em pressupostos:

> Scientific observations, theories and concepts are always mediated by presumptions, and competing views and concepts exist in almost every field of knowledge. The most important task of research on KOS [knowledge organization systems] is to argue which conceptions should be preferred as the basis on which to construe and evaluate KOS. (HJØRLAND, 2008, p. 1528–1529)

No capítulo seguinte, abordaremos os recursos terminológicos produzidos e estudados pela OC, os quais têm sido designados por diversos termos, tais como linguagens documentais, vocabulários controlados e SOC.

2 Recursos terminológicos em OC: linguagens, vocabulários e sistemas

Na literatura em ciência da informação, verifica-se uma proliferação de termos genéricos para designar os recursos tradicionalmente utilizados em bibliotecas e outras unidades de informação para a OC, desde *linguagem documental* a *sistema de organização do conhecimento*. Podemos agrupar estas designações da seguinte forma:

- *linguagem documental, documentária* ou *de indexação;*
- *vocabulário controlado* ou *estruturado;*
- *sistema de organização do conhecimento.*

Propomos, na próxima secção, uma breve análise dos conceitos subjacentes a estas designações em OC.

2.1 Linguagens documentais, documentárias ou de indexação

O primeiro conjunto de termos listado acima inclui as designações mais antigas nos textos de especialidade. O uso dos equivalentes em inglês (*documentary language*) e francês (*langage documentaire*) é atestado nas normas internacionais em informação e documentação, pelo menos desde a década de 1980[1]. *Linguagem* possui aqui o mesmo significado relativamente aos termos complexos *linguagem artificial* ou *linguagem de programação*, ou seja, é entendida como sistema formal e explícito de signos.

[1] De acordo com a informação recolhida, a redação da antiga norma ISO 5127-6:1983 – *Documentation and information – vocabulary – Part 6: Documentary languages*, terá começado em 1979. Disponível em: https://www.iso.org/ Acesso em: 06 jul. 2021.

Para Campos (2001, p. 17), as linguagens documentais são definidas através da sua função como "instrumentos utilizados para representar o conhecimento de uma dada área do saber". Esta autora observa a base formal e conceptual destas linguagens, nas quais os conceitos são identificados por símbolos, de forma a permitir a sua manipulação:

> Os conceitos, para serem manipulados, necessitam de um símbolo que permita a comunicação. Na área da documentação, o símbolo é lingüístico, sendo denominado "termo de recuperação". Os conceitos e termos são, portanto, elementos de qualquer esquema de classificação e dos tesauros. (CAMPOS, 2001, p. 17)

Esta formulação também está patente nas normas portuguesas e internacionais, em que *linguagem documental* é definida como "linguagem formal utilizada para caracterizar os dados ou o conteúdo de documentos e permitir o seu armazenamento e recuperação" (NP 4285-4, 2000, §4.1.1-01). Na mesma linha, o termo *indexing language* é definido como "artificial language established to characterize the content or form of a document" (ISO 5127, 2017, §3.8.1.06). Ao contrário de uma língua natural, uma linguagem formal ou artificial caracteriza-se por ter regras explícitas, conhecidas *a priori*, no que diz respeito ao seu léxico (ou vocabulário), sintaxe e semântica.

No caso das linguagens documentais com base nas línguas naturais, tais como os tesauros e as listas de cabeçalhos, o *controlo* ou *normalização* do vocabulário é um elemento-chave. O manual SIPORbase (Sistema de Indexação em Português), muito utilizado nas bibliotecas portuguesas, consiste precisamente num conjunto de regras para a normalização dos termos: "O vocabulário documental é controlado, face à linguagem natural, em dois aspectos essenciais: a forma dos termos e o seu significado." (ÁREA DE CLASSIFICAÇÃO E INDEXAÇÃO DA BIBLIOTECA NACIONAL, 1998, f. 3, p. 4). Neste particular, o controlo do vocabulário inclui a limitação do número de conceitos e termos, a desambiguação entre homónimos (através da adição de qualificadores parentéticos)[2], a limitação do âmbito de aplicação dos termos (através de notas de âmbito)[3] e o controlo da sinonímia (distinguindo entre descritores e não-descritores).

2.2 Vocabulários controlados e vocabulários estruturados

De acordo com Harpring (2010), os vocabulários controlados são ferramentas que promovem a consistência da indexação de documentos em catálogos documentais, através da distinção entre termos preferenciais e variantes. Por outro lado, os vocabulários controlados

[2] Os qualificadores parentéticos são adicionados aos termos para fins de desambiguação (ISO 25964-1, 2011, §6.2.2). Por exemplo, podemos querer distinguir entre os conceitos de teatro enquanto instituição e teatro enquanto edifício: "teatros (instituições)" e "teatros (edifícios)".

[3] As notas de âmbito têm como função clarificar o uso de um conceito de um vocabulário controlado no âmbito da indexação por assuntos (ISO 25964-1, 2011, §5.2). Por exemplo, o conceito de "iluminuras" num vocabulário controlado poderá incluir a seguinte nota de âmbito: "Abrange tanto a decoração ornamental como as ilustrações em manuscritos e incunábulos, desde que feitas à mão".

facilitam a recuperação da informação, pois levam os utilizadores dos catálogos a utilizar os mesmos termos atribuídos pelos indexadores:

> A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain. (HARPRING, 2010, p. 12)

Como a autora refere, os vocabulários controlados são muitas vezes também *vocabulários estruturados*, no sentido em que dão ênfase às relações entre os conceitos denotados pelos termos. Neste sentido, as atuais normas ISO em informação e documentação tornam mais explícita a dimensão conceptual dos vocabulários controlados. As normas especificam também que um vocabulário controlado não contém necessariamente termos das línguas naturais, podendo consistir em códigos alfanuméricos ou outro tipo de símbolos, com regras próprias para a sua elaboração. Assim, *controlled vocabulary* é definido nas normas como "prescribed list of terms, headings or codes, each representing a concept" (ISO 25964-1, 2011, §2.12), enquanto que *structured vocabulary* é entendido como "organized set of terms, headings or codes representing concepts and their inter-relationships, which can be used to support information retrieval" (ISO 25964-1, 2011, §2.56).

2.3. Sistemas de organização do conhecimento (SOC)

A designação *sistema de organização do conhecimento*, e seus equivalentes, é mais recente nos textos de especialidade. No entanto, o conceito de SOC não surge nas normas internacionais em informação e documentação, tais como a ISO 5127 (2017) e a ISO 25964-1 (2011), nas quais os equivalentes ingleses de *linguagem de indexação* e *vocabulário controlado* são utilizados. O conceito de SOC é, apesar disso, predominante na literatura científica em OC, uma vez que engloba não só as linguagens documentais utilizadas em bibliotecas e outras unidades de informação, mas também os modelos formais para representação do conhecimento na *web*, como é o caso das ontologias.

Hodge (2000) avançou com uma das primeiras tentativas de definição do conceito de SOC, abrangendo as linguagens documentais, vocabulários controlados e os esquemas de conceitos da *web* semântica:

> The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, and authority files that control variant versions of key information (such as geographic names and personal names). They also include less-traditional schemes, such as semantic networks and ontologies. (HODGE, 2000, p. 3)

Moreiro González (2011) salienta que todos os recursos designados por *linguagens documentais* são SOC, embora variem em termos de estrutura e composição, desde listas de

palavras-chave até recursos baseados em conceitos e relações conceptuais, tais como os tesauros e as ontologias. Por outro lado, como salienta o mesmo autor, nem todos os SOC podem ser caracterizados como vocabulários controlados, apesar da ênfase das tradicionais linguagens documentais no controlo terminológico. As chamadas *folksonomias*, isto é, conjuntos de palavras-chave livremente atribuídas pelos utilizadores de um catálogo, portal ou outro tipo de sítio *web*, são exemplos de SOC em que não se verifica controlo de vocabulário (MOREIRO GONZÁLEZ, 2011).

Como podemos verificar, o conceito de SOC permanece algo difuso, abrangendo diversos tipos de recursos terminológicos. Veremos, em seguida, algumas das suas tipologias mais relevantes na literatura de especialidade.

2.4 Tipologias de SOC

A tabela 1 apresenta a tipologia de SOC proposta por Hodge (2000), a qual serviu de base para as propostas subsequentes de diversos autores. Hodge agrupa os SOC em três categorias empíricas: (i) listas de termos, (ii) classificações e categorias e (iii) listas de relações.

Tabela 1: Tipologia de SOC de acordo com Hodge (2000)

Categorias	***Tipos de SOC***
Listas de termos	Ficheiros de autoridade
	Índices toponímicos
	Glossários
	Dicionários
Classificações e categorias	Listas de cabeçalhos de assunto
	Esquemas de classificação
	Taxonomias
Listas de relações	Tesauros
	Redes semânticas
	Ontologias

Fonte: elaborada pelo autor.

Segundo esta proposta, as listas de termos incluem os ficheiros de autoridade, ou seja, listas para controlo das designações autorizadas e variantes para pessoas, organizações, nomes geográficos e outro tipo de entidades. Hodge (2000) salienta, todavia, que os ficheiros de autoridade podem estar organizados de acordo com um esquema de classificação e podem mesmo conter alguma estrutura hierárquica. A autora acrescenta ainda, nas listas de termos, os índices toponímicos (*gazetteers*, em inglês), em que cada entrada poderá incluir o tipo de lugar

(p. ex., país, rio, edifício) e coordenadas geográficas. Finalmente, Hodge (2000) integra nesta categoria os glossários e dicionários, embora não clarifique a utilização destes recursos lexicográficos na organização e recuperação de informação.

As classificações e categorias incluem, de acordo com Hodge (2000), as listas de cabeçalhos de assunto, que consistem em listas de descritores para os assuntos dos documentos numa coleção, juntamente com regras para a sua combinação. Desta categoria, fazem ainda parte os esquemas de classificação, os quais permitem reunir documentos nas mesmas classes e subclasses, e as taxonomias, que consistem numa estrutura hierárquica de classes, sendo muito usadas em empresas no contexto de sistemas de gestão do conhecimento.

Hodge (2000) introduz nas chamadas *listas de relações* os tesauros, responsáveis pela organização dos conceitos num domínio, através de relações hierárquicas, associativas e instanciais, apresentando os termos preferenciais e não preferenciais para cada conceito. As redes semânticas, por seu turno, tendem a dar mais ênfase às relações não hierárquicas entre conceitos, incluindo diversos tipos destas relações (por exemplo, causa-efeito, processo-agente). Por último, as ontologias constituem o tipo de SOC mais recente, sendo definidas como modelos de conhecimento num qualquer domínio. Enquanto modelos formais, as ontologias baseiam-se em axiomas, ou declarações formais, para descrição das classes, propriedades e indivíduos num domínio do conhecimento.

Outra tipologia influente foi proposta por Zeng (2008). Esta tipologia é baseada na estrutura e função dos SOC, consistindo nas seguintes categorias: (i) listas de termos, (ii) modelos semelhantes a metadados, (iii) classificações e categorizações e (iv) modelos relacionais.

Na sua generalidade, as listas de termos e os modelos semelhantes a metadados são planos, embora os índices toponímicos, diretórios e ficheiros de autoridade possam ter alguma estrutura. As classificações e categorizações apresentam uma estrutura bidimensional, através de relações hierárquicas entre as classes, categorias ou assuntos. Finalmente, os modelos relacionais caracterizam-se pela sua complexidade, podendo apresentar diversos tipos de relações não hierárquicas.

Zeng (2008) baseia a sua tipologia nas seguintes funções: (i) eliminar a ambiguidade, (ii) controlo dos sinónimos, (iii) estabelecer relações hierárquicas entre conceitos, (iv) estabelecer relações associativas entre conceitos, e (v) explicitar as propriedades ou atributos.

Como testemunho da natureza difusa do conceito, Souza, Tudhope e Almeida (2012) incluem ainda como categorias de SOC o texto não estruturado (por exemplo, os resumos), as linhas de concordâncias da linguística de *corpus* e os modelos e diagramas conceptuais. Esta perspetiva, no entanto, vai além da relação com a terminologia, inserindo-se no sentido mais lato da OC enquanto área de estudos (HJØRLAND, 2008).

Tabela 2: Tipologia de SOC de acordo com Zeng (2008)

Categorias	***Tipos de SOC***
Listas de termos	Listas de palavras-chave
	Dicionários
	Glossários
	Anéis de sinónimos
Modelos semelhantes a metadados	Ficheiros de autoridade
	Diretórios
	Índices toponímicos
Classificações e categorizações	Listas de cabeçalhos de assunto
	Esquemas de categorização
	Taxonomias
	Esquemas de classificação
Modelos relacionais	Tesauros
	Redes semânticas
	Ontologias

Fonte: elaborada pelo autor.

Como referido por Bratková e Kučerová (2014), a proposta de Zeng (2008) que descrevemos acima foi posteriormente revista e transformada no *KOS Types Vocabulary*[4], que fornece os tipos de SOC a serem indicados no NKOS (*Networked Knowledge Organization Systems*)[5], um perfil de metadados para descrição de SOC na *web*. Uma das alterações mais significativas consistiu na adição das terminologias, entendidas como conjuntos de conceitos e designações em domínios de especialidade ou assuntos (ISO 1087, 2019, §3.1.11).

Os principais tipos de SOC listados no *KOS Types Vocabulary* serviram também como base para a norma ISO relativa à interoperabilidade dos tesauros com outros tipos de SOC (designados aqui por *vocabulários*), incluindo apenas os esquemas de classificação, as taxonomias, os cabeçalhos de assunto, as ontologias, as terminologias, os ficheiros de autoridades de nomes e os anéis de sinónimos (ISO 25964-2, 2013). Esta tipologia mais restrita integra apenas os SOC normalmente utilizados para a recuperação de informação, ou que possam ser reutilizados para a construção de tesauros. Por exemplo, as terminologias podem constituir-se como fontes de vocabulário para a construção de tesauros com vista à recuperação de informação.

[4] Disponível em https://nkos.slis.kent.edu/nkos-type.html Acesso em: 06 jul. 2021.

[5] Disponível em: https://nkos.slis.kent.edu/nkos-ap.html Acesso em: 06 jul. 2021.

3 SKOS (Simple Knowledge Organization System): modelização de SOC na web

O SKOS (*Simple Knowledge Organization System*) é uma recomendação do W3C (*World Wide Web Consortium*) para a modelização de SOC através de tecnologias da *web* semântica[6]. Desenvolvido no seguimento de vários projetos europeus desde o final da década de 1990, o SKOS foi concebido para a migração de SOC para a *web*, nomeadamente através da sua representação em RDF (*Resource Description Framework*)[7], um dos fundamentos da *web* semântica (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

A motivação para o desenvolvimento do SKOS é elucidativa do fosso entre as tradicionais linguagens documentais e os modelos formais para representação do conhecimento. Como referem Baker *et al.* (2013), instrumentos como os tesauros e os esquemas de classificação constituem SOC *informais*, ao contrário das ontologias. Como tal, é inviável a tradução direta destes sistemas para axiomas de classes e propriedades:

> Informally defined KOSs cannot typically be translated into the language of RDFS and OWL properties and classes, with their formal-logical implications, without introducing potentially false or misleading logical precision. Informal KOSs may be converted into formal ontologies […], but the process of assigning appropriate formal semantics to the elements of a KOS may require a long, hard modeling effort. (BAKER *et al.*, 2013, p. 38)

Os autores avançam com o exemplo das relações hierárquicas, que num tesauro, abrangem tipicamente as relações genéricas, partitivas e de instanciação (ISO 25964-1, 2011). A tradução destas relações para uma ontologia requer a desambiguação entre relações de subsunção (classe-subclasse), de classe-indivíduo, parte-todo e possivelmente outras, dada a natureza imprecisa destes recursos. Por outro lado, os tradicionais vocabulários controlados mantêm-se relevantes hoje, com aplicações que vão desde a melhoria da precisão na recuperação de informação à expansão da pesquisa. Desta forma, o SKOS permite uma transição de baixo custo de SOC informais para a *web* semântica (BAKER *et al.*, 2013).

O modelo de dados do SKOS (Fig. 1) permite representar os SOC como esquemas de conceitos. O SKOS define *conceito* enquanto unidade de pensamento, ecoando as definições presentes em versões mais antigas das normas ISO em terminologia. No entanto, esta definição é apenas sugestiva, uma vez que deve abranger a organização intelectual da generalidade dos SOC, embora se aplique mais diretamente aos tesauros e esquemas de classificação enquanto exemplos prototípicos da aplicação do SKOS.

Os conceitos podem se relacionar entre si através de relações semânticas, constituindo hierarquias informais (*broader*, *narrower*) ou redes de associações (*related*). Por outro lado, os

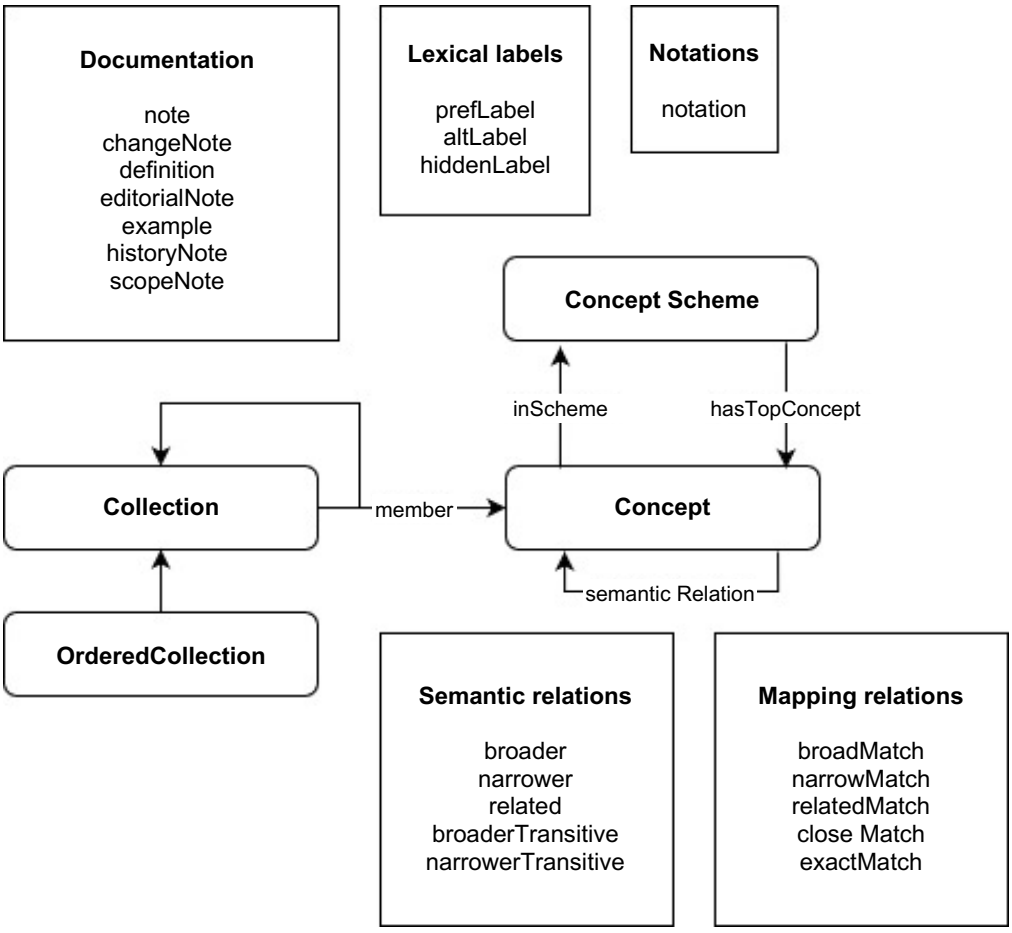
[6] *SKOS Simple Knowledge Organization System Reference.* Disponível em: https://www.w3.org/TR/skos-reference/ Acesso em: 06 jul. 2021.

[7] *Resource Description Framework.* Disponível em: https://www.w3.org/RDF/ Acesso em: 06 jul. 2021.

conceitos podem ser incluídos em coleções não hierárquicas (simples ou ordenadas), sendo designados por formas lexicais (*lexical labels*) – preferenciais, alternativas e rejeitadas – e identificadas por notações alfanuméricas.

Existem também propriedades de documentação que permitem associar diversos tipos de notas, exemplos e definições. As propriedades de mapeamento permitem estabelecer relações com conceitos de esquemas externos, o que é essencial para a publicação de SOC como dados ligados (*linked data*)[8].

Figura 1: Modelo de dados do SKOS (baseado em Baker *et al.*, 2013)



Fonte: elaborada pelo autor.

Uma das limitações do SKOS consiste na representação dos termos, uma vez que o SKOS apenas permite que os conceitos sejam designados por *etiquetas lexicais*, as quais correspondem linguisticamente a formas de termos simples, compostos ou complexos. Tal implica que as formas singular e plural do mesmo termo (por exemplo, *gato* e *gatos*) possam designar o mesmo conceito num esquema conceptual, enquanto etiquetas preferencial e

[8] *Linked Data*. Disponível em: https://www.w3.org/wiki/LinkedData Acesso em: 06 jul. 2021.

alternativa. Por outro lado, o SKOS não permite representar qualquer tipo de informação associada às formas lexicais, como é o caso das relações lexicais ou semânticas entre termos da mesma língua.

Para permitir uma maior flexibilidade na modelização das formas lexicais, foi introduzida a extensão SKOS-XL (*SKOS eXtension for Labels*)[9], que inclui a classe *Label* para as formas lexicais, assim como uma propriedade genérica para estabelecer relações ao nível das formas lexicais, podendo esta propriedade ser especializada através da definição de subpropriedades para representar relações lexicais específicas. Isto permite, por exemplo, representar a relação entre as formas *Países Africanos de Língua Oficial Portuguesa* e o seu acrónimo *PALOP*, o que não seria possível no modelo SKOS simples.

3.1 O SKOS do ponto de vista da bidimensionalidade da terminologia

A crescente aproximação entre as normas internacionais em terminologia e em informação e documentação traduz-se também ao nível dos modelos para representação de SOC, como é o caso do SKOS. Tendo em conta que as terminologias podem ser incluídas nesta categoria de recursos, ainda que com finalidades e características diferentes, importa avaliar a capacidade do SKOS em modelizar terminologias, entendidas como conjuntos de conceitos e termos em domínios ou assuntos especializados.

Tomando a norma ISO 1087 (2019) como referência, verificamos que um recurso terminológico é constituído por um conjunto de *entradas terminológicas*, as quais reúnem dados sobre um só conceito e as suas designações, sejam termos em línguas naturais, nomes próprios ou símbolos. A informação a apresentar sobre os conceitos numa entrada terminológica deverá incluir *relações conceptuais* (hierárquicas ou associativas), *designações*, tais como termos, símbolos ou nomes próprios (no caso dos chamados *conceitos individuais*) e *definições*. Poderá ainda integrar informação sobre os termos *preferenciais*, *admitidos*, *rejeitados* ou *obsoletos* para um mesmo conceito, *contextos* de uso e relações léxico-semânticas (sinonímia, equivalência e antonímia).

A tabela 3 apresenta correspondências possíveis entre os elementos da ISO 1087 acima indicados e o vocabulário do SKOS. Aí verificamos que o SKOS permite modelizar a grande maioria dos elementos presentes na norma sobre terminologia, embora alguns elementos (como as relações léxico-semânticas) exijam a definição de propriedades através da extensão SKOS-XL[10].

[9] *Appendix B. SKOS eXtension for Labels (SKOS-XL)*. Disponível em: https://www.w3.org/TR/skos-reference/#xl Acesso em: 06 jul. 2021.

[10] Existe também uma limitação no que diz respeito à indicação de símbolos não linguísticos como designações. Embora seja possível fazê-lo através da propriedade *notation* do SKOS, dois conceitos não podem ter a mesma notação.

Tabela 3: Correspondência entre elementos ISO 1087 e SKOS

Elemento ISO 1087	**Elemento SKOS**
Recurso terminológico	*Concept scheme*
Conceito (geral ou individual)	*Concept*
Relação hierárquica (genérica, partiviva)	*Broader, narrower*
Relação associativa (sequencial, espacial, temporal, causal)	*Related*
Critério de subdivisão	*Collection*
Definição (por intensão, por extensão)	*Definition*
Designação (termo, nome próprio, símbolo)	*Lexical label, notation*
Termo preferencial	*Preferred label*
Termo admitido	*Alternative label*
Termo rejeitado	*Hidden label*
Termo obsoleto	[Não existe]
Contexto	*Example*
Sinonímia	[Através do SKOS-XL]
Equivalência	[Através do SKOS-XL]
Antonímia	[Através do SKOS-XL]

Fonte: elaborada pelo autor.

Uma outra questão pertinente prende-se com o controlo do vocabulário. O SKOS permite o controlo do vocabulário através da indicação de uma forma preferencial para cada conceito, a qual não pode ser associada a mais nenhum conceito. No entanto, a atribuição de etiquetas preferenciais não é obrigatória no SKOS, o que possibilita a modelização de recursos terminológicos sem controlo de vocabulário. Neste caso, os conceitos seriam representados apenas por etiquetas alternativas (que podem ser atribuídas a mais do que um conceito).

Em suma, verificamos que o SKOS estabelece a distinção entre as dimensões linguística e conceptual do trabalho terminológico, sendo flexível o suficiente para a modelização de recursos terminológicos baseados no conhecimento. A extensão SKOS-XL permite ainda modelizar relações léxico-semânticas entre designações.

Conclusão

Neste artigo, partimos do pressuposto que as ferramentas utilizadas na organização do conhecimento, desde tesauros a esquemas de classificação, podem ser entendidas como recursos terminológicos. Começamos por abordar as relações teóricas e metodológicas entre a terminologia e a organização do conhecimento, áreas de saber marcadas pela interdisciplinaridade.

Assumindo como enquadramento teórico uma abordagem bidimensional, baseada na distinção entre a dimensão linguística e a dimensão conceptual da terminologia, propomos uma breve análise dos conceitos subjacentes às designações *linguagem documental*, *vocabulário controlado* e *sistema de organização do conhecimento*. Verificamos que os conceitos subjacentes não são idênticos, embora as suas designações sejam por vezes utilizadas de forma indiscriminada: existem SOC onde não se verifica controlo de vocabulário (por exemplo, as chamadas *folksonomias*) e que não podem ser caracterizadas como linguagens documentais (por exemplo, as ontologias). Este desfasamento ao nível das designações é indicativo da grande disparidade de tipologias de SOC nos textos de especialidade e reflete uma conceção mais abrangente da OC enquanto disciplina.

A nossa análise termina com a descrição do SKOS (*Simple Knowledge Organization System*), um modelo para a representação de SOC na *web* semântica. Avaliamos o modelo segundo a abordagem bidimensional da terminologia, aferindo a sua capacidade para representar os principais elementos das entradas terminológicas. Segundo a nossa análise, o SKOS, incluindo a sua extensão para as etiquetas lexicais, permite modelizar a grande maioria dos elementos presentes na norma ISO 1087, em que se definem os fundamentos e vocabulário da terminologia enquanto disciplina.

Este artigo vem, desta forma, confirmar a crescente aproximação teórica e metodológica entre a terminologia e a OC. A aproximação entre as referidas áreas manifesta-se também nas normas internacionais em ciência da informação, em que foram transpostos conceitos basilares da terminologia. Por exemplo, na norma ISO 5127, as definições de *objeto*, *conceito* e *característica*, entre outras, têm como fonte a ISO 1087. A aproximação entre terminologia e OC leva a que as ferramentas da OC, desde as linguagens documentais aos mais recentes SOC da *web* semântica, possam ser abordadas como recursos terminológicos de pleno direito. As implicações desta abordagem são significativas para a terminologia, que pode tomar partido de modelos como o SKOS para a representação de recursos terminológicos na *web* semântica, de forma interoperável com os SOC. Por outro lado, esta abordagem tem vantagens para a própria OC, onde podem ser aplicados enquadramentos teóricos nascidos no seio da terminologia, como é o caso da abordagem bidimensional, para a construção e análise de SOC. Tal perspetiva implica o reconhecimento das dimensões linguística e conceptual da terminologia, abrindo a possibilidade de aplicar métodos semasiológicos, baseados na análise de *corpora* linguísticos, em paralelo a métodos onomasiológicos, focados na análise do conhecimento em domínios especializados.

Referências

ALMEIDA, B.; COSTA, R. OntoAndalus: an ontology of Islamic artefacts for terminological purposes. *Semantic Web Journal,* v. 12, n. 2, p. 295–311, 2021.

ALMEIDA, B.; COSTA, R.; ROCHE, C. The names of lighting artefacts: extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus. *Revue TAL*, v. 60, n. 3, p. 113–137, 2019.

ALMEIDA, B.; FREIRE, N.; MONTEIRO, D. The development of the ROSSIO Thesaurus: supporting content discovery and management in a research infrastructure. DOSSO, D.; FERILLI, S.; MANGHI, P.; POGGI, A.; SERRA, G.; SILVELLO, G. In: (Eds.). *Proceedings of the 17th Italian Research Conference on Digital Libraries*. Aachen: CEUR-WS, 2021. Disponível em: http://ceur-ws.org/Vol-2816/ Acesso em: 06 jul. 2021.

ÁREA DE CLASSIFICAÇÃO E INDEXAÇÃO DA BIBLIOTECA NACIONAL. *SIPORbase: Sistema de Indexação em Português : manual*. 3a ed. rev. e aumentada. Lisboa: Biblioteca Nacional, 1998.

BAKER, T.; BECHHOFER, S.; ISAAC, A.; MILES, A.; SCHREIBER, G.; SUMMERS, S. Key Choices in the Design of Simple Knowledge Organization System (SKOS). *Journal of Web Semantics*, v. 20, p. 35–49, maio 2013.

BOWKER, L.; PEARSON, J. *Working with specialized language : a practical guide to using corpora*. London: Routledge, 2002.

BRATKOVÁ, E.; KUČEROVÁ, H. Knowledge Organization Systems and Their Typology. *Revue of Librarianship*, v. 25, n. 2, p. 1–25, 2014.

CABRÉ, M. T. Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles*, n. 21, p. 10–15, 2000.

CABRÉ, M. T. Theories of terminology: their description, prescription and explanation. *Terminology*, v. 9, n. 2, p. 163–199, 2003.

CABRÉ, M. T. La teoría comunicativa de la terminología: una aproximación lingüística a los términos. *Revue française de linguistique appliquée*, v. 14, n. 2, p. 9–15, 2009.

CAMPOS, M. L. A. *Linguagem documentária: teorias que fundamentam sua elaboração*. Niterói: Universidade Federal Fluminense, 2001.

COSTA, R. Plurality of theoretical approaches to terminology. In: PICHT, H. (Ed.). *Modern approaches to terminological theories and applications*. Bern: Peter Lang, 2006. p. 79–89.

COSTA, R. Terminology and specialised lexicography: two complementary domains. *Lexicographica*, v. 29, n. 1, p. 29–42, 2013.

COSTA, R.; SILVA, R.; CAMPOS, M. I. B. Terminologia, uma disciplina de interfaces. *Linha d'Água*, v. 33, n. 1, p. 1–8, abr. 2020.

DAHLBERG, I. A referent-oriented, analytical concept theory for INTERCONCEPT. *International Classification*, v. 5, n. 3, p. 142–151, 1978.

DAHLBERG, I. Knowledge organization and terminology: philosophical and linguistic bases. *International Classification*, v. 19, n. 2, p. 65–71, 1992.

DAHLBERG, I. Brief communication: Concepts and terms - ISKO's major challenge. *Knowledge Organization*, v. 36, n. 2/3, p. 169–177, 2009.

FABER, P. (Ed.). *A cognitive linguistics view of terminology and specialized language.* Berlin: De Gruyter Mouton, 2012.

FELBER, H. *Terminology manual.* Paris: UNESCO, 1984.

FILLMORE, C. J. Frames and the semantics of understanding. *Quaderni di semantica*, v. 6, n. 2, p. 222–254, dez. 1985.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: STAAB, S.; STUDER, R. (Eds.). *Handbook on ontologies*. Second ed. Berlin: Springer, 2009. p. 1–17.

HARPRING, P. *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Los Angeles: Getty Research Institute, 2010.

HITZLER, P.; KRÖTZSCH, M.; RUDOLPH, S. *Foundations of semantic web technologies*. Boca Raton: CRC Press, 2010.

HJØRLAND, B. What is Knowledge Organization (KO)? *Knowledge Organization*, v. 35, n. 2/3, p. 86–101, 2008.

HJØRLAND, B. Concept theory. *Journal of the American Society for Information Science and Technology*, v. 60, n. 8, p. 1519–1536, 2009.

HODGE, G. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. [S.l.] Council on Library and Information Resources, 2000.

ISO 704. *Terminology work – Principles and methods*. Geneva: ISO, 2009.

ISO 1087. *Terminology work and terminology science – Vocabulary.* Geneva: ISO, 2019.

ISO 5127. *Information and documentation - Foundation and vocabulary.* Geneva: ISO, 2017.

ISO 25964-1. *Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval.* Geneva: ISO, 2011.

ISO 25964-2. *Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies.* Geneva: ISO, 2013.

LAKOFF, G. *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press, 1987.

LAKOFF, G.; JOHNSON, M. *Metaphors we live by.* Chicago: University of Chicago Press, 1980.

MELBY, A. K. Terminology in the age of multilingual corpora. *The Journal of Specialised Translation*, n. 18, p. 7–29, jul. 2012.

MOREIRO GONZÁLEZ, J. A. *Linguagens documentárias e vocabulários semânticos para a web: elementos conceituais*. Salvador: EDUFBA, 2011.

NP 4285-4. *Documentação e informação - Vocabulário - Parte 4: Linguagens documentais.* Caparica: IPQ, 2000.

NUOPPONEN, A. Tangled web of concept relations: concept relations for ISO 1087-1 and ISO 704. *TKE 2014: Ontology, Terminology & Text Mining*, n. p., jun. 2014.

ROCHE, C. *Terme et concept : fondements pour une ontoterminologie.* TOTh 2007. Annecy: Institut Porphyre, 2007.

SAGER, J. C. *A practical course in terminology processing.* Amsterdam: John Benjamins, 1990.

SANTOS, C.; COSTA, R. Domain specificity: semasiological and onomasiological knowledge representation. In: KOCKAERT, H. J.; STEURS, F. (Eds.). *Handbook of terminology*. Amsterdam: John Benjamins, 2015. v. 1, p. 153–179.

SILVA, R. *Gestão de terminologia pela qualidade: processos de validação*. Tese de doutoramento - Universidade NOVA de Lisboa, Lisboa, 2014.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. Towards a taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. *Knowledge Organization*, v. 39, n. 3, p. 179–192, jan. 2012.

TEMMERMAN, R. *Towards new ways of terminology description: the sociocognitive approach.* Amsterdam: John Benjamins, 2000.

WIERZBICKA, A. *Semantics: primes and universals.* Oxford: Oxford University Press, 1996.

WÜSTER, E. *Introduction to the general theory of terminology and terminological lexicography*. Vienna: Springer, 1979.

ZENG, M. Knowledge Organization Systems (KOS). *Knowledge Organization,* v. 35, n. 2/3, p. 160–182, 2008.

Recebido: 07/07/2021.

Aprovado: 23/07/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 47-64, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p47-64

Artículo / Artigo / Article

Documentación de lenguas amenazadas en la época digital

Documentação de línguas ameaçadas na era digital

Endangered Language Documentation in the Digital Age

Mika Hämäläinen*
mika.hamalainen@helsinki.fi
https://orcid.org/0000-0001-9315-1278

Jack Rueter**
jack.rueter@helsinki.fi
https://orcid.org/0000-0002-3076-7929

Khalid Alnajjar***
khalid.alnajjar@helsinki.fi
https://orcid.org/0000-0002-7986-2994

Resumen

Presentamos nuestra infraestructura para la documentación de lenguas urálicas, que consiste en herramientas para redactar diccionarios de tal forma que las entradas sean estructuradas en el formato XML (Extensible Markup Language). Desde los diccionarios en XML podemos generar código para analizadores morfológicos que son útiles para todo tipo de actividades de PLN. En este artículo mostramos las ventajas que una documentación digital y legible por máquina tiene. Describimos, también, el sistema en el contexto de lenguas urálicas amenazadas.

Palavras-chave: Diccionarios digitales; Procesamiento de lenguajes naturales; Lenguas urálicas; Documentación lingüística; Infraestructura abierta.

* Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

** Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

*** Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

Resumo

Apresentamos a nossa infraestrutura para documentação de línguas urálicas, a qual consiste num conjunto de ferramentas para redigir dicionários de modo que as entradas sejam estruturadas num formato XML (Extensible Markup Language). A partir de dicionários XML, podemos criar códigos para analisadores morfológicos, que são úteis a todos os tipos de atividades de processamento de língua natural. Neste artigo, demonstramos as vantagens da documentação digital legível por máquina e descrevemos o sistema no contexto das línguas urálicas ameaçadas.

Palavras-chave: *Dicionários digitais; Processamento de linguagens naturais; Línguas urálicas; Documentação linguística; Infraestrutura aberta.*

Abstract

We present our infrastructure to document Uralic languages, which consists of tools to write dictionaries so that entries are structured in XML (Extensible Markup Language) format. From dictionaries in XML, we can generate code for morphological analysers useful for all kinds of NLP tasks. In this article, we show the advantages of digital and machine-readable documentation. We also describe the system in the context of endangered Uralic languages.

Keywords: *Digital Dictionaries; Natural Language Processing; Uralic Languages; Linguistic Documentation; Open Infrastructure.*

Introducción

La mayoría de las lenguas habladas en el mundo están amenazadas y están en peligro de extinción. Su documentación y revitalización son de altísimo valor cultural, por lo cual han recibido mucha atención académica en varias disciplinas como en la antropología, tipología, lexicografía y lingüística computacional. No obstante, los recursos producidos en cada proyecto no serán necesariamente publicados de forma abierta ni para la comunidad de los hablantes nativos ni para el uso de otros proyectos científicos.

El objetivo de nuestro artículo es describir nuestra abierta infraestructura para documentar lenguas minoritarias. Presentamos nuestras experiencias con las siguientes lenguas urálicas: sami de skolt (sms), erzya (myv), moksha (mdf), komi ziriano (kpv) y komi permio (koi). Como pertenecen a la rama urálica, son lenguas que exhiben una amplia riqueza morfológica lo que hace su tratamiento automático un desafío para los métodos modernos apoyados en el aprendizaje automático. Sin embargo, realizando la documentación lingüística de forma estructurada que permite la lectura automática, es posible crear los recursos computacionales necesarios para el PLN (procesamiento de lenguajes naturales) al mismo tiempo con la documentación lingüística.

Estamos a punto de comenzar a trabajar con la lengua apurinã (apu), lo que nos permite reflejar nuestro contexto urálico desde una perspectiva más amplia, y aumenta la relevancia de

nuestro trabajo en el contexto latinoamericano. Describimos, así, cómo nuestra infraestructura puede funcionar en contextos fuera de lo urálico.

La documentación lingüística es un campo de estudio académico que se ha desarrollado considerablemente en las últimas décadas. Su objetivo es proporcionar un registro completo de las prácticas lingüísticas características de una comunidad de habla determinada (HIMMELMANN, 1998). El objetivo de la documentación del lenguaje es crear el registro más completo posible de la comunidad de habla tanto para las futuras generaciones como para la revitalización del idioma. El resultado de dicho trabajo se manifiesta como un *corpus* lingüístico u otro tipo de colección de material. Estos datos son una documentación del idioma que puede analizarse y estudiarse de diversas maneras. La cuestión si los materiales actualmente documentados realmente describen el uso del lenguaje de una comunidad de habla con éxito puede ser discutible, y al menos este objetivo nunca podrá alcanzarse por completo. Sin embargo, especialmente en la época actual con lenguas en peligro de extinción, estos materiales, a menudo, constituyen los únicos recursos disponibles en estos idiomas.

Si los materiales de documentación lingüística deberían ser accesibles y cómo deberían ser distribuidos, ha sido un tema de debate. Creemos que es importante entender que esto también es una cuestión de granularidad, y la pregunta no es necesariamente si los materiales son accesibles, sino más bien qué partes deberían permitir qué tipo de acceso. Existen razones buenas para mantener materiales culturalmente sensibles disponibles solo para los grupos específicos. Al mismo tiempo, siempre hay materiales en cualquier idioma que son más neutrales y que los propios autores quieren hacer accesibles. Especialmente para trabajos escritos y publicados, siempre puede ser posible negociar una publicación con licencias abiertas modernas, lo que también permitiría la reutilización de los mismos materiales en diferentes propósitos de investigación abierta.

Estos materiales son particularmente importantes cuando desarrollamos herramientas de PLN, porque este trabajo puede beneficiarse mucho de los recursos más transparentes y accesibles que posible. En las secciones siguientes analizaremos ejemplos de dicho trabajo, incluido el contexto de los treebanks de dependencias universales. Hay que enfatizar que la tecnología abierta desarrollada en una infraestructura abierta también puede usarse para procesar materiales que están disponibles solo para un investigador en particular o miembros individuales de la comunidad. Por lo tanto, la infraestructura abierta beneficia tanto a los entornos de uso abiertos como cerrados, mientras que una infraestructura cerrada, posiblemente, solo beneficia a los grandes actores comerciales.

1 Estudios relacionados

Hay varios proyectos individuales en diferentes partes del mundo que trabajan con diccionarios en línea para lenguas amenazadas. Sin embargo, muchos proyectos tienen una lengua en su enfoque y trabajan sin conocer otros proyectos con otras lenguas amenazadas. Esto

ha llegado en una situación en que los investigadores resuelven el mismo tipo de problemas individualmente para su lengua de interés. En este apartado, presentamos algunos proyectos digitales.

El trabajo con idiomas en peligro de extinción en América del Norte ha demostrado que se debe proporcionar herramientas de aprendizaje al principiante en un idioma. Las comunidades son pequeñas y la falta de familiaridad con la tradición lexicográfica puede fácilmente ser perjudicial para la experiencia de aprendizaje del principiante. No se puede esperar que el estudiante de un nuevo idioma sepa dónde se encuentra una entrada del diccionario ni que adopte automáticamente la ortografía normativa. Cuando el usuario del idioma carece del teclado o del conocimiento para escribir correctamente, las estrategias de relajación ortográfica se pueden implementar en soluciones que dominan la morfología para móviles y en línea. El conocimiento morfológico y la relajación de ortografía se utilizan para atender a los principiantes en lenguas tsimshiánicas y salishanas en el uso de diccionarios y el PLN (LITTELL et al., 2017).

En un frente completamente separado, también se ha trabajado para proveer a la comunidad de Yupik de la isla de St. Lawrence el acceso sin obstáculos a materiales lingüísticos en línea. Esto ha sido posible utilizando un diccionario morfológicamente consciente. En su sistema, se ha introducido una estrategia de múltiples métodos de entrada que atienden a diferentes sistemas de escritura (HUNT et al., 2019). El trabajo aquí está hecho a medida, y se mantiene un fuerte vínculo entre un idioma y su comunidad. Estos idiomas en peligro de extinción se incluyen en la categoría de lenguas de bajos recursos.

Lo problemático es que "lengua de bajos recursos" es un término que se utiliza para casi cualquier idioma con menor presencia en Internet que el inglés. Lenguas como el hindi (IRVINE y CALLISON-BURCH, 2014), el árabe (CHEN et al., 2018) o bien el persa (AHMADNIA et al., 2017) son consideradas a menudo lenguas de bajos recursos en el mundo de PLN, aunque tienen millones de hablantes. En el trabajo de Nasution et al., (2018), por lo contrario, las lenguas malasias son relativamente pequeñas en comparación con las lenguas mayoritarias que las rodean. El enfoque consiste en trabajar simultáneamente con un grupo de idiomas muy relacionados en una infraestructura multilingüe e independiente del idioma. Los autores analizan el uso de las entradas de un diccionario bilingüe y explican la dificultad de seleccionar los diccionarios bilingües adecuados para comenzar la documentación.

Una de las infraestructuras más ambiciosas para la documentación de lenguas minoritarias desde el punto de vista de la lingüística computacional es, sin duda, la de Giella (MOSHAGEN et al., 2014). Su infraestructura está basada en dos componentes principales: los transductores TEF (transductores de estados finitos) y diccionarios en XML. Los transductores son una forma de documentar la morfología de una lengua de manera computacional. Es decir, son colecciones de reglas sobre cómo el sistema morfológico de una lengua funciona. Estas reglas pueden usarse directamente para un análisis automático de texto y para conjugar lemas en sus variantes morfológicas.

Se utilizan los transductores y los diccionarios para herramientas como corrección ortográfica en *Word*[1], predicción de texto en teclados de *Android* e iOS[2], sistemas interactivos para aprender lenguas (BONTOGON et al., 2018) y diccionarios en línea (RUETER y HÄMÄLÄINEN, 2017). Nuestra infraestructura está basada en Giella, lo que nos permite sincronizar los datos entre las dos infraestructuras. Esto significa que los avances en la documentación lingüística en nuestra infraestructura pueden usarse directamente en las herramientas producidas en Giella.

2 Nuestra infraestructura para la documentación lingüística

Giella requiere una competencia relativamente alta en programación para poder redactar diccionarios y programar reglas morfológicas en los transductores, y al mismo tiempo, requiere buenos conocimientos en la lengua que está en curso de la documentación. Su infraestructura es demasiado complicada incluso para los que han estudiado informática, y por lo tanto no es accesible para una comunidad fuera de los que colaboran directamente con Giella. Por este motivo, nuestra infraestructura tiene varias interfaces para distintos tipos de usuarios; tanto para usuarios que no tienen conocimientos suficientes para escribir XML o programar transductores como para desarrolladores que quieren utilizar las herramientas sin saber cómo compilarlas desde cero.

2.1 Diccionarios XML en Línea

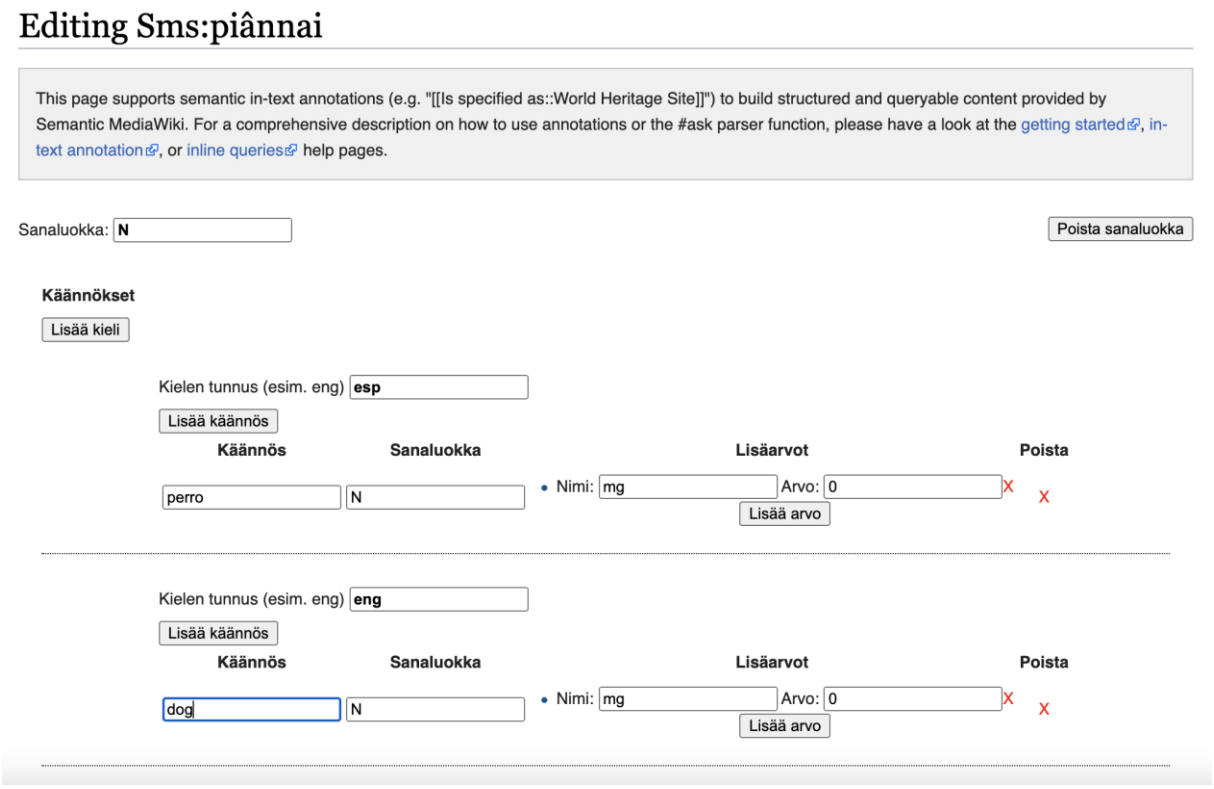
Un paso muy importante en la documentación de una lengua minoritaria es el trabajo lexicográfico. Esto resulta en un diccionario que puede ser útil tanto para los hablantes nativos como para los que quieran aprender el idioma. Nosotros guardamos los diccionarios en el formato XML muy estructurado. Eso quiere decir que todo tipo de metadatos están en sus propios campos en vez de estar guardados en varios partes de una entrada lexicográfica de forma no estructurada. Esto es importante ya que no solo queremos guardar los diccionarios para el uso de un ser humano, sino también queremos que sean legibles de forma automática.

Nuestro sistema *Akusanat*[3] (HÄMÄLÄINEN y RUETER, 2019a) está basado en *MediaWiki* y permite visualizar el contenido de los diccionarios XML para todo tipo de usuarios. Los datos del *MediaWiki* están sincronizados con los archivos XML utilizando el control de versiones *Git*. Esto significa que, si alguien modifica una entrada lexicográfica en Akusanat, estos cambios resultarán en un cambio en el diccionario XML almacenado en *GitHub*. Si alguien modifica los diccionarios XML directamente, Akusanat descargará los

[1] Disponible en: http://divvun.no/korrektur/korrektur.html. Accedido en: 11 jul 2021
[2] Disponible en: http://divvun.no/keyboards/mobileindex.html. Accedido en: 11 jul 2021
[3] Disponible en: https://akusanat.com. Accedido en: 11 jul 2021

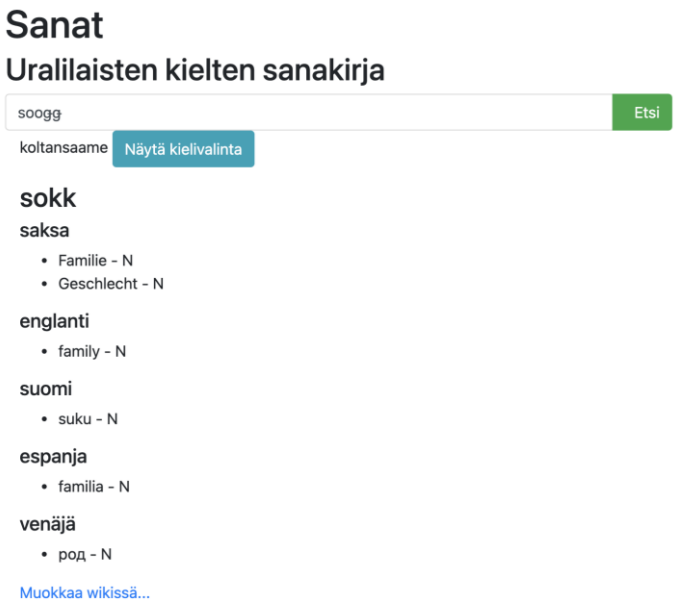
nuevos cambios desde *GitHub* y actualizará su base de datos de forma automática. Esto hace posible que usuarios avanzados puedan editar los archivos XML directamente con su herramienta favorita y que los usuarios menos avanzados puedan hacer cambios en línea con una interfaz. Akusanat no deja a los usuarios modificar la sintaxis *Wiki* directamente, sino que muestra un formulario que asegura que los cambios siguen siendo estructurados y compatibles con XML (Fig. 1).

Figura 1: El formulario en Akusanat para editar la entrada *piânnai* (perro) en sami de skolt



Para la búsqueda utilizamos los transductores para procesar el input del usuario. Esto quiere decir que el usuario puede buscar una palabra en cualquiera de sus conjugaciones morfológicas, ya que el TEF puede lematizar palabras de forma automática. También es posible buscar palabras escritas de manera errónea. Los transductores contienen información sobre los errores de ortografía más comunes en cada lengua, lo que nos permite resolver el lema, aunque la palabra no haya sido escrita según la norma ortográfica. Esto es importante en el caso de las lenguas con las que trabajamos, ya que las normas ortográficas no son tan establecidas como en el caso de lenguas mayoritarias.

Figura 2: La interfaz para realizar búsquedas en Akusanat



La Figura 2 muestra la interfaz para buscar palabras en el diccionario. En el ejemplo, el término de búsqueda es la palabra en sami de skolt *soogg* que es el genitivo de la palabra *sokk* que significa *familia*. Nuestro sistema lematiza el término de búsqueda automáticamente con el TEF de sami de skolt, y muestra la entrada para el lema *sokk* al usuario.

La idea de utilizar *MediaWiki*, y sobre todo *Semantic MediaWiki*, para crear diccionarios no es nueva, ya que ya existen varios proyectos que utilizan la tecnología como su base (MULJADI et al., 2006; BON y NOWAK, 2013; DUEÑAS y GÓMEZ, 2015). Aunque, sin duda, *MediaWiki* tiene sus ventajas, en práctica nosotros hemos tenido que programar nuestras propias extensiones *MediaWiki* para añadir la funcionalidad necesaria; el formulario para editar, la sincronización de *MediaWiki*-XML, la búsqueda con los transductores etc. El problema que hemos experimentado muchas veces es que el funcionamiento interno de *MediaWiki* cambia demasiado a menudo. Esto significa que, si queremos mantener el *MediaWiki* actualizado con las últimas actualizaciones de seguridad, tenemos que hacer muchos cambios en nuestro código fuente para mantener el funcionamiento de nuestras extensiones con la nueva versión de *MediaWiki*. Aun así, seguimos utilizando y desarrollando Akusanat[4] por el momento, ya que ofrece un entorno sencillo para los usuarios. En el siguiente apartado, describimos el otro sistema que estamos desarrollando, y que, algún día, podrá sustituir Akusanat.

2.2 Redacción de Diccionarios

En este apartado, describimos el sistema Ve′rdd[5] (ALNAJJAR et al., 2020). El sistema funciona con los mismos diccionarios XML que Akusanat y puede usarse en línea de la misma

[4] El código fuente está disponible en https://github.com/mikahama/akusanat. Accedido en: 11 jul 2021

[5] Disponible en: https://akusanat.com/verdd/. Accedido en: 11 jul 2021

manera. La diferencia está en el enfoque del sistema. Ve′rdd no es un sistema para visualizar las entradas lexicográficas para un usuario final, sino un sistema creado específicamente para redactar diccionarios tanto digitales como impresos. Para realizar el sistema, hemos colaborado con un grupo de profesionales que trabajan con diccionarios impresos.

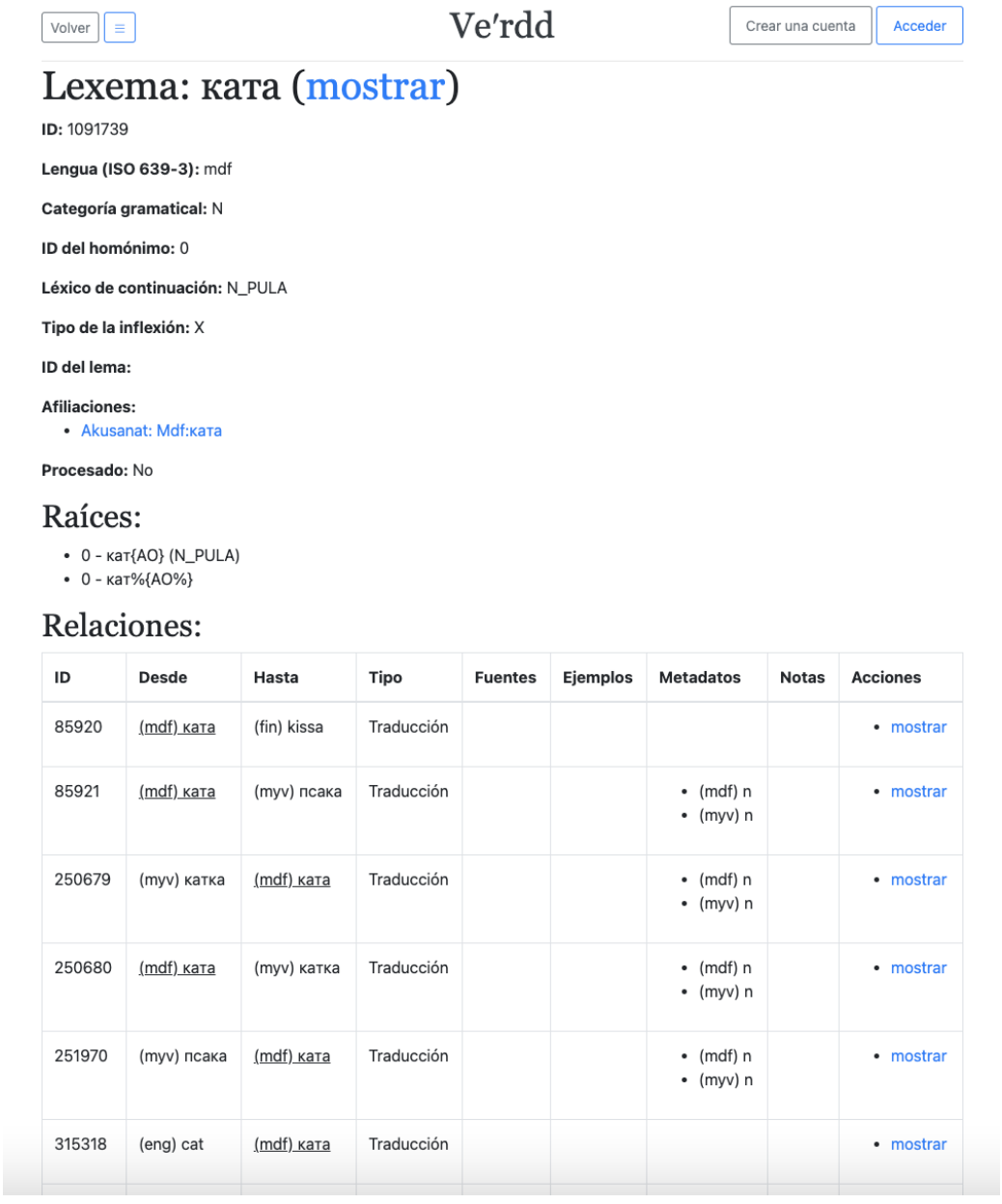
Con las lenguas con las que trabajamos, la documentación lexicográfica no empieza desde cero, ya que tanto las lenguas sami habladas en los países nórdicos como las lenguas pérmicas y mordvínicas habladas en Rusia han recibido mucha atención por su documentación durante el siglo pasado. Por ejemplo, para el sami de skolt existe el diccionario de Sammallahti & Mosnikoff (1991), y existen varios estudios sobre las lenguas mordvínicas (AASMÄE et al., 2016; GRÜNTHAL, 2016) y pérmicas (HAMARI, 2011; KLUMPP, 2016). Si hay diccionarios en forma digital, existen en un formato sin estructurar como un archivo *Word*, CSV o bien PDF producido con un sistema de ROC (reconocimiento óptico de caracteres). Por este motivo, Ve′rdd incluye funcionalidad para importar datos lexicográficos de formatos sin estructurar. Hemos prestado mucha atención en la calidad de la conversión, ya que, en el caso de nuestras lenguas, sobre todo, en el caso de sami de skolt, es muy frecuente que exista el mismo carácter con muchas codificaciones diferentes. Por ejemplo, ′ (U+02B9 modificador de letra prime) es un carácter muy común en sami de skolt, pero por la razón del teclado finlandés, es a menudo escrito como ' (U+0027 apóstrofo) o bien ´ (U+00B4 acento agudo). Ve′rdd está programado para tomar en cuenta los caracteres posibles de la lengua e intentar a corregir los caracteres erróneos automáticamente.

Figura 3: La interfaz para realizar búsquedas y filtrar entradas léxicas en Ve′rdd

La Figura 3 muestra la interfaz para realizar búsquedas y filtrar palabras en Ve′rdd. La interfaz está diseñada para apoyar el flujo de trabajo del editor del diccionario. Por ejemplo, es

posible mostrar solamente las entradas sin procesar. Esto significa entradas que nadie ha verificado después de importar los datos desde un formato sin estructurar. Para facilitar el desarrollo de los transductores es también posible ordenar y filtrar las palabras según el léxico de continuación. El léxico de continuación es una forma de expresar que una palabra se conjuga del mismo modo que otras palabras con el mismo léxico de continuación.

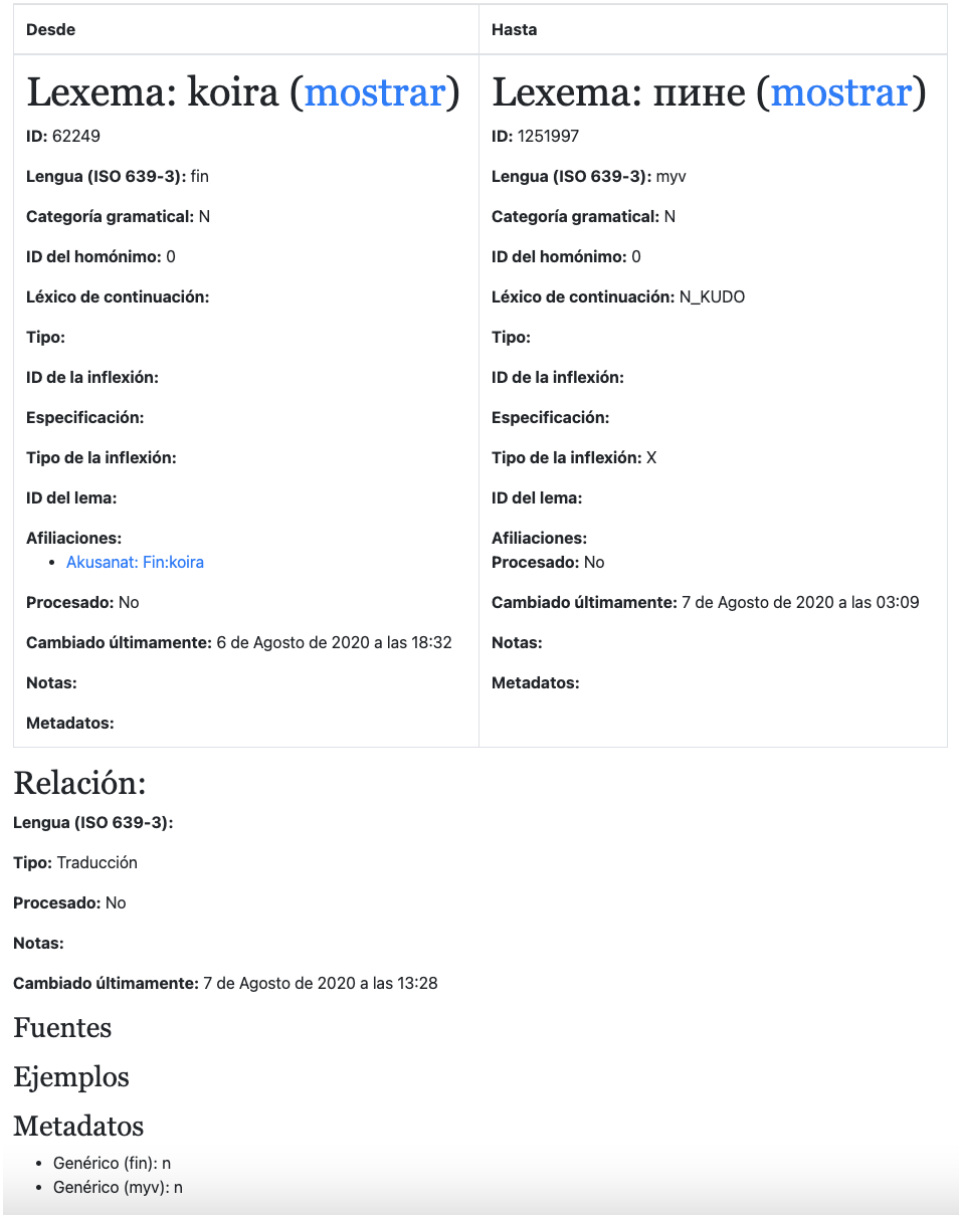
Figura 4: La interfaz para editar entradas léxicas en Ve′rdd



Aparte de solamente buscar y filtrar entradas léxicas, es importante tener la posibilidad de editarlas. La Figura 4 muestra la interfaz para inspeccionar una entrada en el diccionario. Si el usuario está conectado con su cuenta, además de ver, puede editar la información de la entrada. Ve′rdd está diseñado para ser una herramienta para diccionarios multilingües, por eso

una entrada está conectada a otras entradas en el sistema. En la figura, se pueden ver relaciones de tipo traducción que conectan una palabra a sus traducciones en otras lenguas. También es posible definir otro tipo de relaciones entre lenguas como relaciones etimológicas. Las relaciones pueden existir entre las palabras de la misma lengua, por ejemplo, es posible indicar palabras compuestas o bien derivaciones con las relaciones. Como los transductores contienen información derivativa, Ve′rdd añade automáticamente este tipo de relaciones al importar un diccionario sin estructurar.

Figura 5: La interfaz para comparar dos entradas relacionadas en Ve′rdd



Ve′rdd puede visualizar la relación entre dos palabras enlazadas con algún tipo de relación para verificar que una palabra en una lengua está enlazada al homónimo correcto en otra lengua (Figura 5). También es posible editar el tipo de la relación o bien borrar las relaciones innecesarias.

En todo momento, Ve′rdd tiene la posibilidad de exportar el diccionario en formatos distintos. Los más importantes para nosotros son el XML de Giella que puede utilizarse para generar los transductores y el código Latex. Desde el código Latex, es posible generar un PDF para imprimir el diccionario. El formato Latex hace posible cambiar el estilo del diccionario sin cambiar el contenido, si hay cambios en Ve′rdd, es posible actualizar el contenido del diccionario sin cambiar el estilo definido en Latex. Esta funcionalidad ha sido un objetivo importante para nosotros ya que el trabajo hecho en Ve′rdd no debería únicamente servirse para lo digital sin que también para editar diccionarios imprimidos.

2.3 Recursos para el PLN

Nuestros sistemas para editar diccionarios son directamente útiles para el desarrollo de los transductores ya que podemos exportar el léxico en el formato necesario para HFST (LINDÉN et al., 2013). HFST es la herramienta que utilizamos para crear los transductores. Nosotros disponemos de transductores para el sami de skolt (RUETER y HÄMÄLÄINEN, 2020), erzya y moksha (RUETER et al., 2020) y las lenguas komi. Los transductores pueden utilizarse para lematizar palabras, analizar su morfología o bien generar formas conjugadas. Estos transductores son difíciles de compilar para personas que no trabajan con los transductores a menudo. Por este motivo, nosotros compilamos todos los transductores cada noche y los distribuimos mediante nuestra página web[6]. No sólo compilamos nuestros transductores sino todos los transductores para todas las lenguas en la infraestructura Giella.

Sin embargo, queda difícil usar los transductores como tal. Por este motivo, hemos desarrollado una librería *Python* llamada UralicNLP (HÄMÄLÄINEN, 2019). Con la librería, es posible descargar transductores y diccionarios compilados, y usarlos directamente en *Python*. Fig 6 muestra cómo utilizar nuestros transductores desde *Python*. En la segunda línea de código, se analiza la palabra *шляпа* (sombrero) en erzya (myv). El resultado indica que la palabra es un nombre (+N) indefinido (+Indef) en el singular (+Sg) del nominativo (+Nom). En la cuarta línea generamos la forma conjugada de la misma palabra en plural (+Pl). El resultado es la palabra plural *шляпат*.

Figura 6: Un ejemplo del uso de UralicNLP

```
>>> from uralicNLP import uralicApi
>>> uralicApi.analyze("шляпа", "myv")
[('шляпа+N+Sg+Nom+Indef', 0.0)]
>>> uralicApi.generate("шляпа+N+Pl+Nom+Indef", "myv")
[('шляпат', 0.0)]
>>>
```

[6] Disponible en: https://models.uralicnlp.com/nightly/. Accedido en: 11 jul 2021

Los transductores producen todas las interpretaciones posibles de una palabra. En el caso de las lenguas urálicas, existe mucha homonimia en la conjugación. Esto significa que, si utilizamos los transductores con un texto corriente, no podemos lematizar las palabras en su contexto ya que los transductores producen todos los lemas posibles. Por este motivo, utilizamos desambiguadores con gramática de restricciones (KARLSSON, 1990) basados en la herramienta llamada VISL CG-3 (BICK y DIDRIKSEN, 2015). Las reglas de gramática de restricciones eliminan morfologías que no son posibles en la oración, y resultan en una oración morfológicamente desambiguada.

Figura 7: Un ejemplo del uso del desambiguador de komi-ziriano

```
>>> from uralicNLP.cg3 import Cg3
>>> oracion = "Ныв ёртыслы гижис письмö"
>>> cg = Cg3("kpv")
>>> print(cg.disambiguate(oracion.split(" ")))
Warning: Line 6 had empty tag.
[('Ныв', [<ныв – N, Sg, Nom, <W:0.000000>>]), ('ёртыслы', [<ёрт – N, Sg, Dat, Px
Sg3, So/PC, <W:0.000000>>]), ('гижис', [<гижны – V, TV, Ind, Prt1, Sg3, <W:0.000
000>>]), ('письмö', [<письмö – N, Sg, Nom, <W:0.000000>>])]
>>>
```

En la Figura 7, podemos ver cómo se utilizan los desambiguadores de gramática de restricciones mediante UralicNLP. En la tercera línea se inicializa el objeto de desambiguación para el komi-ziriano (kpv) y en la cuarta línea se llama el método de desambiguación del objeto con una oración. El resultado contiene la forma de la palabra en la oración, su lematización y su morfología para cada palabra de la oración.

Aparte de los diccionarios estructurados y las herramientas basadas en reglas, disponemos de treebanks de las dependencias universales para el sami de skolt, moksha, erzya (RUETER y TYERS, 2018), komi-ziriano (PARTANEN et al., 2018) y komi-permio (RUETER et al., 2020). Estos treebanks contienen anotaciones sintácticas con las etiquetas morfológicas de las dependencias universales. Con los últimos treebanks, también hemos añadido las etiquetas morfológicas que producen los transductores para facilitar el uso de los dos recursos juntos.

3 Los Beneficios de la Documentación Digital

La documentación digital nos ha permitido utilizar los últimos métodos en el mundo de PLN para aumentar automáticamente los datos que tenemos en los diccionarios. Cómo todos los diccionarios XML con los que trabajamos son multilingües, el primer paso que hemos dado con la tecnología de PLN ha sido la predicción de traducciones (HÄMÄLÄINEN et al., 2018). La idea ha sido la siguiente: si el diccionario de sami de skolt contiene traducciones al finés, alemán e inglés, y el diccionario de erzya contiene traducciones al finés, inglés, ruso y francés, entonces, con esta información, debería ser posible deducir automáticamente traducciones de sami de skolt al ruso y francés y de erzya al alemán dada la existencia de dos lenguas en común:

finés e inglés. Con un modelo probabilístico hemos aumentado la cantidad de traducciones en los diccionarios de sami de skolt, erzya, moksha y komi-ziriano. Después de obtener los resultados automáticos, hemos comprobado las traducciones de forma manual antes de incluirlas en los diccionarios XML.

Como las redes neuronales exigen una gran cantidad de datos para ser entrenados, es habitual creer que su uso no es posible en el caso de las lenguas en peligro. Nosotros hemos tomado la perspectiva que podemos generar la cantidad de datos necesaria para una red neuronal con nuestras herramientas morfológicas. Utilizando los treebanks y los transductores, hemos generado datos para entrenar una red neuronal para realizar desambiguación en vez de utilizar la gramática de restricciones para erzya y komi-ziriano (ENS et al., 2019). La idea ha sido generar todos los análisis posibles para las palabras en los treebanks y entrenar la red neuronal para desambiguar los análisis con el análisis del treebank. También hemos podido utilizar las redes neuronales para aumentar las relaciones etimológicas en el diccionario de sami de skolt (HÄMÄLÄINEN y RUETER, 2019b).

Nuestras herramientas son compatibles con la infraestructura de Giella. Esto ha hecho posible utilizar nuestros diccionarios y transductores directamente en su plataforma en línea para aprender idiomas (ANTONSEN y ARGESE, 2018), en los teclados de *Android* y *iPhone* y en la corrección ortográfica para *Word* y *OpenOffice* desarrollados por Divvun[7] en colaboración con Giella.

La documentación digital tiene claramente sus beneficios, ya que podemos realizar aprendizaje automático con diccionarios estructurados y transductores de morfología. Por este motivo el proyecto conducido en la universidad de Oulu para redactar el nuevo diccionario finés-sami de skolt ha optado por utilizar Ve′rdd para crear el diccionario. Hemos trabajado juntos con los empleados del proyecto para aumentar la funcionalidad de nuestro sistema. Ve′rdd ha hecho posible el trabajo simultáneo de los editores que, sin Ve′rdd, hubiesen utilizado Excel y Word para su trabajo. Esto habría significado una posibilidad pérdida de producir un diccionario estructurado para el interés de PLN y un diccionario impreso al mismo tiempo.

Como hemos desarrollado todos los recursos y herramientas de forma abierta, incluso investigadores ajenos han empezado a utilizar los recursos mediante UralicNLP. La librería de *Python* ha sido utilizada, entre otros, por Creutz y Sjöblom (2019) para corregir textos escritos por hablantes no-nativos. Avikainen (2019) ha utilizado la librería para investigar noticias en periódicos antiguos digitalizados y Rämö (2020) para generar títulos de noticias de forma automática.

[7] Disponible en: http://divvun.no/. Accedido en: 11 jul 2021

Conclusiones y el trabajo futuro

En este artículo hemos presentado nuestras soluciones para la documentación digital de lenguas urálicas amenazadas. Como trabajamos con muchas lenguas a la vez en colaboración con otro proyecto de infraestructura para lenguas minoritarias, Giella, podemos diseñar nuestras herramientas e infraestructura de tal modo que comenzar el trabajo con una nueva lengua no requiere que inventemos la rueda de nuevo, sino que podemos incorporar la lengua fácilmente en todas las herramientas compatibles. Esto quiere decir, que con el trabajo lexicográfico en Ve'rdd, podemos crear transductores de forma fácil. Y, al tener un transductor, los recursos de la lengua ya pueden utilizarse para tareas más complejas como teclados, corrección ortográfica y todo tipo de tareas de PLN mediante UralicNLP.

A la hora de escribir este artículo, hemos comenzado a trabajar con la lengua apurinã. Gracias a los esfuerzos de su documentación lingüística (FACUNDES, 2000), podemos expresar sus reglas en el formalismo de TEF. Hemos comenzado a incorporar sus materiales lexicográficos en nuestra infraestructura. Aunque el trabajo aún está en sus fases iniciales, las experiencias por ahora han sido positivas. La morfología de apurinã es muy distinta a la morfología urálica, pero la robustez de la tecnología de TEF nos permite modelar su morfología.

Estamos muy interesados en colaborar con la documentación digital de todo tipo de lenguas amenazadas. Desde el punto de vista del PLN, los recursos multilingües son más útiles para todo tipo de tareas que los recursos mono- o bilingües. Como ya hemos visto en este artículo, si tenemos diccionarios multilingües en el mismo sistema, podemos aumentar la cantidad de traducciones que tienen de una forma automática.

Todas las herramientas y recursos que hemos descrito en este artículo están disponibles de forma abierta en *GitHub*[8]. En nuestro trabajo, siempre utilizamos licencias abiertas y almacenamos datos de forma permanente en Zenodo.

Referencias

AASMÄE, N.; PAJUSALU, K.; KABAJEVA, N. Gemination in the Mordvin Languages. *Linguistica Uralica*, 52(2). 2006. Disponible en: https://www.ceeol.com/search/article-detail?id=396961. Accedido en: 11 jul 2021

AHMADNIA, B.; SERRANO, J.; HAFFARI, G. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. En *Proceedings of RANLP*. 2017 p. 24-30. DOI: 10.26615/978-954-452-049-6_004. Accedido en: 11 jul 2021

ALNAJJAR, K.; HÄMÄLÄINEN, M.; RUETER, J.; PARTANEN, N. Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement. En *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. 2020 p. 1-6. DOI: 10.18653/v1/2020.coling-demos.1. Accedido en: 11 jul 2021

[8] Los enlaces están disponibles en https://uralicnlp.com/. Accedido en: 11 jul 2021

ANTONSEN, L.; ARGESE, C. Using authentic texts for grammar exercises for a minority language. En *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings. 2018 p. 1–9. Disponible en: https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=152&Article_No=1 Accedido en: 11 jul 2021

AVIKAINEN, J. *A Method for Wavelet-Based Time Series Analysis of Historical Newspapers*. Universidad de Helsinki. Tesina de Master. 2019. Disponible en: https://helda.helsinki.fi/handle/10138/310021. Accedido en: 11 jul 2021

BICK, E.; DIDRIKSEN, T. Cg-3—beyond classical constraint grammar. En *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. 2015. p. 31-39. Disponible en: https://aclanthology.org/W15-1807. Accedido en: 11 jul 2021

BON, B.; NOWAK, K. Wiki lexicographica. Linking medieval latin dictionaries with semantic MediaWiki. En *Electronic lexicography in the 21st century: thinking outside the paper:* proceedings of the eLex 2013 conference, Estonia, 2013, p. 407-420. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=4565204. Accedido en: 11 jul 2021

BONTOGON, M.; ARPPE, A.; ANTONSEN, L.; THUNDER, D.; LACHLER, J. Intelligent Computer Assisted Language Learning (ICALL) for nêhiyawêwin: An In-Depth User-Experience Evaluation. En *Canadian Modern Language Review,* 74(3). 2018. p. 337-362. DOI: https://doi.org/10.3138/cmlr.4054. Accedido en: 11 jul 2021

CHEN, X.; SUN, Y.; ATHIWARATKUN, B.; CARDIE, C.; WEINBERGER, K. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6. 2018. p. 557-570. Disponible en: https://arxiv.org/abs/1606.01614. Accedido en : 11 jul 2021

CREUTZ, M.; SJÖBLOM, E. E. Toward automatic improvement of language produced by non-native language learners. En *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. 2019. p. 20-30. Disponible en: https://aclanthology.org/W19-6303. Accedido en: 11 jul 2021

DUEÑAS, G.; GÓMEZ, D. A bilingual dictionary with Semantic Mediawiki: The language Saliba's case. En The *4th International Conference on Language Documentation and Conservation (ICLDC)*. 2015. Disponible en: http://hdl.handle.net/10125/25338. Accedido en : 11 jul 2021

ENS, J.; HÄMÄLÄINEN, M.; RUETER, J.; PASQUIER, P. Morphosyntactic Disambiguation in an Endangered Language Setting. En *22nd Nordic Conference on Computational Linguistics (NoDaLiDa): Proceedings of the Conference.* 2019. p. 345-349. Disponible en: https://aclanthology.org/W19-6139. Accedido en: 11 jul 2021

FACUNDES, S. D. S. The language of the Apurinã people of Brazil. Buffalo: State University of New York at Buffalo (Dissertation). 2000. Disponible en: http://www.etnolinguistica.org/tese:facundes-2000. Accedido en: 11 jul 2021

GRÜNTHAL, R. Transitivity in Erzya: Second language speakers in a grammatical focus. En *Mordvin languages in the field*. Finno-Ugrian Society. 2016. p. 291-318. Disponible en:

https://researchportal.helsinki.fi/en/publications/transitivity-in-erzya-second-language-speakers-in-a-grammatical-f. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M. UralicNLP: An NLP Library for Uralic Languages. *Journal of open source software*, 4(37). 2019. Disponible en: https://joss.theoj.org/papers/10.21105/joss.01345. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; RUETER, J. An open online dictionary for endangered Uralic languages. En *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, 111. 2019a. Disponible en: http://hdl.handle.net/10138/305873. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; RUETER, J. Finding Sami Cognates with a Character-Based NMT Approach. En *Proceedings of the 3rd Workshop on Computational Methods in the Study of Endangered Languages: (Volume 1) Papers.* 2019b. p. 39-45. Disponible en: https://aclanthology.org/W19-6006. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; TARVAINEN, L. L.; RUETER, J. Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018 p. 862-867. Disponible en: https://aclanthology.org/L18-1138. Accedido en: 11 jul 2021

HAMARI, A. The abessive in the Permic languages. En *Suomalais-Ugrilaisen Seuran Aikakauskirja*, 2011(93). 2011. p.37-84. DOI: https://doi.org/10.33340/susa.82172. Accedido en: 11 jul 2021

HIMMELMANN, N. P. Documentary and descriptive linguistics. *Linguistics*, 36. 1998. p. 161-196. DOI: https://doi.org/10.1515/ling.1998.36.1.161. Accedido en : 11 jul 2021

HUNT, B.; CHEN, E.; SCHREINER, S. L.; SCHWARTZ, L. Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 2019 pp. 122–126. DOI: 10.18653/v1/N19-4021. Accedido en: 11 jul 2021

IRVINE, A.; CALLISON-BURCH, C. Hallucinating phrase translations for low resource mt. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning.* 2014. p. 160-170. DOI: 10.3115/v1/W14-1617. Accedido en: 11 jul 2021

KARLSSON, F. Constraint Grammar as a Framework for Parsing Unrestricted Text. En *Proceedings of the 13th International Conference of Computational Linguistics, Vol. 3*. 1990. p. 168-173. DOI: https://doi.org/10.3115/991146.991176. Accedido en : 11 jul 2021

KLUMPP, G. Semantic functions of complementizers in Permic languages. En *Complementizer Semantics in European Languages*, 2016. p. 529-586. DOI: https://doi.org/10.1515/9783110416619-016. Accedido en: 11 jul 2021

LINDÉN, K.; AXELSON, E.; DROBAC, S.; HARDWICK, S.; KUOKKALA, J.; NIEMI, J.; PIRINEN, T.; SILFVERBERG, M. HFST - A System for Creating NLP Tools. En *Systems and Frameworks for Computational Morphology. Communications in Computer and Information Science*. 380. Humboldt-Universität in Berlin: Springer. 2013. p. 53-71. DOI: 10.1007/978-3-642-40486-3_4. Accedido en: 11 jul 2021

LITTELL, P.; PINE, A.; DAVIS, H. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. En *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics. 2017. p. 141–150. DOI: 10.18653/v1/W17-0119. Accedido en: 11 jul 2021

MOSHAGEN, S.; RUETER, J.; PIRINEN, T.; TROSTERUD, T.; TYERS, F. M. Open-source infrastructures for collaborative work on under-resourced languages. En *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*. 2014. p. 71-77. Disponible en: http://www.syros.aegean.gr/users/spyrosv/papers/ccurl14.pdf#page=78. Accedido en: 11 jul 2021

MULJADI, H.; TAKEDA, H.; KAWAMOTO, S.; KOBAYASHI, S.; FUJIYAMA, A. Towards a Semantic Wiki-Based Japanese Biodictionary. En *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics*. 2006. Disponible en: http://www-kasm.nii.ac.jp/papers/takeda/06/muljadi06eswc.pdf. Accedido en: 11 jul 2021

NASUTION, A.H.; MURAKAMI, Y.; ISHIDA, T. Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. En *Proceedings of the 11th Language Resources and Evaluation Conference. European Language Resource Association.* 2018. Disponible en: https://aclanthology.org/L18-1536. Accedido en: 11 jul 2021

PARTANEN, N.; BLOKLAND, R.; LIM, K.; POIBEAU, T.; RIESSLER, M. The first Komi-Zyrian universal dependencies treebanks. En *Second Workshop on Universal Dependencies (UDW 2018)*. 2018. p. 126-132. DOI: 10.18653/v1/W18-6015. Accedido en: 11 jul 2021

RÄMÖ, M. (Re)lexicalization of auto-written news with contextual and cross-lingual word embeddings. Universidad de Helsinki. Tesina de Master. 2020. Disponible en: https://helda.helsinki.fi/handle/10138/321924. Accedido en: 11 jul 2021

RUETER, J.; HÄMÄLÄINEN, M. FST Morphology for the Endangered Skolt Sami Language. En *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020).* 2020. p. 250-257. Disponible en: https://aclanthology.org/2020.sltu-1.35. Accedido en: 11 jul 2021

RUETER, J.; HÄMÄLÄINEN, M.; PARTANEN, N. Open-Source Morphology for Endangered Mordvinic Languages. En *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS).* The Association for Computational Linguistics. 2020. p. 94–100. DOI: 10.18653/v1/2020.nlposs-1.13. Accedido en: 11 jul 2021

RUETER, J. M.; HÄMÄLÄINEN, M. Synchronized Mediawiki based analyzer dictionary development. En *3rd International Workshop for Computational Linguistics of Uralic Languages Proceedings of the Workshop*. 2017. DOI: 10.18653/v1/W17-0601. Accedido en: 11 jul 2021

RUETER, J.; PARTANEN, N.; PONOMAREVA, L. On the questions in developing computational infrastructure for Komi-Permyak. En *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages.* 2020 p. 15-25. Disponible en: https://aclanthology.org/2020.iwclul-1.3. Accedido en: 11 jul 2021

RUETER, J. M.; TYERS, F. M. Towards an open-source universal-dependency treebank for Erzya. En *Proceedings of International Workshop for Computational Linguistics of Uralic Languages.* 2018. DOI: 10.18653/v1/W18-0210. Accedido en: 11 jul 2021

SAMMALLAHTI, P.; MOSNIKOFF, J. Suomi-Koltansaame sanakirja. LÄÄ′DD-SÄÄ′m SÄÄ′NNÊ′RJJ Ohcejohka: Girjegiisá Oy. 1991.

Recibido: 29/01/2021.

Aceptado: 05/04/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 65-82, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p65-82

Artigo / Article

O ensino da língua egípcia clássica no Brasil: desafios e possibilidades usando recursos digitais

Teaching Ancient Egyptian Language in Brazil: Challenges and Opportunities of Digital Resources

Ronaldo Guilherme Gurgel Pereira*
ronaldo.gurgel@yahoo.de
https://orcid.org/0000-0002-8457-6220

Thais Rocha da Silva**
thais.rochadasilva@hmc.ox.ac.uk
https://orcid.org/0000-0003-0616-1924

Resumo

Este artigo apresenta os resultados iniciais de um projeto mais amplo sobre o ensino da língua egípcia no Brasil por meio de recursos digitais. Examinamos a primeira etapa através do curso Introdução ao Egípcio Clássico (Egípcio Médio), o primeiro curso online de ensino da língua egípcia no Brasil, tendo em vista a desafiadora realidade para a formação de egiptólogos no país. A partir do debate das Humanidades Digitais, que problematiza a produção e a divulgação do conhecimento nas ciências humanas com recursos tecnológicos, apontamos possíveis caminhos para a expansão deste projeto no cenário brasileiro. Neste escopo, a discussão sobre as plataformas digitais e o ensino de história antiga no Brasil entram como elementos importantes na contextualização desta iniciativa.

Palavras-chave: Egiptologia; Humanidades digitais; Hieróglifos; História antiga; Educação.

Abstract

This paper presents the preliminary results of a larger project for the study of the Egyptian language in Brazil using digital resources. We examined the first stage, Introduction to Classical Egyptian (Middle Egyptian), the first online course for the Egyptian language in Brazil, taking into account the challenging reality of the Egyptological training in the country. Based on the Digital Humanities debate that questions the production and

* Centro de Humanidades – CHAM, Faculdade de Ciências Sociais e Humanas, FCSH, Universidade NOVA de Lisboa, Lisboa, Portugal.

** Universidade de São Paulo - USP, São Paulo, Brasil; Faculdade de Filosofia, Letras e Ciências Humanas, São Paulo. Pesquisadora na University of Oxford - Harris Manchester, Oxford, Inglaterra.

dissemination of knowledge in the human sciences using technological resources, we point out possible ways for the expansion of this project in the Brazilian scenario. In this scope, the discussion about digital platforms and the teaching of Ancient History in Brazil are important elements in the context of this initiative.

Keywords: *Egyptology; Digital Humanities; Hieroglyphs; Ancient History; Education.*

Introdução

A Egiptologia, segundo o imaginário popular, é sinônimo de ruínas antigas e solitárias, em meio ao deserto, à espera de arqueólogos para revelarem os seus tesouros e mistérios. Contudo, a transição da Egiptologia de um *hobby* aristocrático para a sua consolidação como campo científico decorre dos esforços de muitos pesquisadores, mas deve a formação da disciplina acadêmica principalmente aos filólogos e linguistas.

Até meados do século XIX dependíamos do relato bíblico ou de viajantes gregos, romanos e árabes para acessarmos o Egito. Ao facultar-se o aprendizado da língua egípcia, os investigadores adquiriram a capacidade de diálogo com aquela sociedade sem passar pelos filtros de relatos estrangeiros. Restituir ao egípcio a capacidade de falar de si depois de um milênio e meio de silêncio se tornou, portanto, o marco fundador da Egiptologia.

O estudo da língua egípcia antiga clássica, o Egípcio Médio, é uma prerrogativa importante para a formação dos egiptólogos em todo o mundo. O conhecimento dos hieróglifos, com a tradução da Pedra de Rosetta em 1822 por Jean-François Champollion, marcou o início da Egiptologia como disciplina acadêmica. Sua origem filológica, no contexto dos estudos orientais na Europa do século XIX, teve desdobramentos variados na implementação do campo ao redor do mundo, que podem ser medidos até hoje nas diversas escolas da Egiptologia e dos seus centros de pesquisa, localizadas principalmente na Europa e na América do Norte.

Como uma disciplina predominantemente francesa, inglesa e alemã, as gramáticas e métodos de aprendizagem da língua egípcia foram produzido nestes idiomas, trazendo muitos obstáculos àqueles que não eram versados nessas línguas. Para os falantes nativos dessas línguas, existe uma rica oferta de bibliografia para se iniciarem e aperfeiçoarem no estudo de qualquer estágio da língua egípcia.

Para os egiptólogos lusófonos[1], o aprendizado das línguas modernas é uma ferramenta indispensável à formação e à continuação das pesquisas, mas nem sempre acessível e realizável em curto prazo. Infelizmente, para o estudante brasileiro, as principais gramáticas e manuais de língua egípcia disponíveis foram produzidos em alemão, francês e inglês.

[1] O mesmo problema é relatado pelos egiptólogos espanhóis. Contudo, a Egiptologia espanhola produziu materiais próprios há algumas décadas, o que certamente favoreceu os estudantes de língua portuguesa, como é o caso da tradução de Collier e Manley (2000). línguaportuguês

As possibilidades de conhecimento e imersão na língua egípcia clássica e suas variantes (neoegípcio, demótico, copta) precisaram enfrentar uma outra barreira linguística que só muito recentemente vem sendo rompida. Em muitos casos, cabia aos aspirantes da Egiptologia, o trabalho de estudar duas línguas simultaneamente: o egípcio antigo e a língua moderna, precisamente o alemão, francês ou inglês, que atuava como mediadora. O processo de aprendizado exigia o dobro do tempo e de esforço, com sucessivas traduções e reformulações dos enunciados e conceitos.

Neste artigo, apresentamos como estudo de caso o curso *Introdução ao Egípcio Clássico (Egípcio Médio)*, uma iniciativa dos autores em parceria com o Grupo de Trabalho de História Antiga da ANPUH (GTHA/ANPUH) e a Universidade Federal de Santa Catarina (UFSC), ministrado entre Setembro e Novembro de 2020. A partir desse curso, exploramos possibilidades que tenham por objetivo amenizar os desafios encontrados pelos pesquisadores e estudantes de Egiptologia no Brasil. A produção de um método em língua portuguesa certamente contribuiu para a barreira linguística (PEREIRA, 2014, 2016).

Contudo, a realidade brasileira ainda possui inúmeros obstáculos para o desenvolvimento dos estudos sobre o Egito antigo. A extensão territorial, o número reduzido de professores especialistas, principalmente com o conhecimento da língua egípcia, a dificuldade de acesso a periódicos especializados, bibliotecas limitadas e a profunda desigualdade social e educacional no país são elementos que precisam ser levados em conta no nosso esforço de formação.

O curso foi o primeiro disponibilizado em plataforma digital com acesso aberto,[2] atingindo todas as regiões do Brasil (e Argentina). Voltado para pesquisadores em diversos estágios de formação e professores da área, a proposta do curso foi também ampliar e consolidar redes de colaboração entre especialistas, favorecendo a troca de informações e materiais e, de maneira secundária, abrir oportunidades para a formação e cultivo de laços de cooperação entre os participantes, incluindo aqueles localizados em estados com menos disponibilidade de professores, livros e grupos de estudo. A transposição de barreiras regionais foi favorecida pelo uso das plataformas digitais, possibilitando que as aulas fossem ministradas de Portugal para o Brasil e a Argentina. Discutiremos o curso à luz dos problemas enfrentados pelo uso das tecnologias e plataformas digitais no ensino, mas buscando apontar caminhos para a ampliação do projeto e de iniciativas semelhantes.

[2] A *playlist* com as 10 aulas introdutórias, que cobrem as formas nominais do Egípcio Clássico, estão disponíveis no canal do YouTube do GTHA/ANPUH e podem ser acessadas pelo link: https://www.youtube.com/playlist?list=PLI8rGh6UbR_vOBaIrALDQwiSHSgQ2TleQ

1 Egiptologia no Brasil: problemas, desafios e oportunidades

A Egiptologia no mundo lusófono ainda é uma disciplina em processo de consolidação e carece de instrumentos essenciais para a capacitação dos seus profissionais. Boas bibliotecas e o treinamento em línguas antigas e modernas são fundamentais para a formação profissional na área, além do acesso às fontes primárias (ROCHA, 2014, 2017, 2019).

No caso específico do Brasil, é preciso levar em conta outros aspectos que não podem ser vistos de forma dissociada. O processo de institucionalização do conhecimento sobre o mundo antigo, a 'chegada' do Egito antigo no Brasil e o desenvolvimento do ensino superior nesta disciplina em território nacional, o qual passa pelo debate sobre a desigualdade social contribuíram para o quadro atual da disciplina em solo brasileiro[3].

O aparecimento tardio da Egiptologia nas universidades não pode ser justificado pela ausência de um interesse pelo Egito antigo ou mesmo pela falta de coleções egípcias no Brasil[4]. O desenvolvimento da chamada História Antiga (onde o Egito se encaixa primordialmente) respondeu a processos específicos no Brasil e que vem sendo avaliados e debatidos na última década (CARVALHO e FUNARI, 2007; GUARINELLO, 2008; SILVA, 2010; SILVA, 2011; FRANCISCO, 2017).

Os textos disponíveis para o ensino da arqueologia, da história e das línguas do mundo antigo estão maioritariamente em língua estrangeira, sobretudo o inglês[5], o que traz dificuldades de todos os tipos aos professores da área que ensinam na graduação. Apesar da existência de traduções de publicações internacionais, elas nem sempre acompanham a velocidade da produção acadêmica e dos debates produzidos pelos centros de excelência, de modo que o material disponível em língua portuguesa fica sempre desatualizado. Desse modo, os professores de história antiga, por vezes injustamente acusados de elitistas por não disponibilizarem textos em língua portuguesa a estudantes, se veem num terreno complicado para ministrarem seus cursos.

No caso do domínio da língua inglesa, um relatório do *British Council* já apontava, em 2014, que, aproximadamente, 5% da população brasileira estava capacitada a interagir com textos complexos em inglês, e que apenas 1% dos brasileiros possuía fluência na língua[6]. De facto, o quadro se mantém, uma vez que o Brasil figura na 53ª posição (*low proficiency*)[7] do relatório de 2020 apresentado pelo *EF English Proficiency Index (EF EPI)* - que analisa dados de 2,3 milhões de falantes não nativos de inglês, em 100 países e regiões[8].

---

3 Sobre a formação da História Antiga no Brasil ver Funari (2010), Santos, Kolv e Nazário (2017), Santos (2019, 2021, no prelo). Sobre a formação da Egiptologia no Brasil, ver Bakos (2004), Rocha (2014, 2017, 2019).

4 Um panorama geral sobre as coleções egípcias no Brasil é apresentado em Bakos (2004) e Brancaglion (2004).

5 O inglês é o idioma mais utilizado nos grupos de pesquisa e nas publicações internacionais.

6 O relatório, preparado pelo *Data Popular Institute,* pode ser consultado na sua íntegra através do link: https://www.britishcouncil.org.br/sites/default/files/learning_english_in_brazil.pdf

7 Por comparação, o outro país lusófono dedicado ao estudo da Egiptologia, Portugal, está colocado na 7ª posição do mesmo ranking (*very high proficiency*).

8 https://www.ef.com/wwen/epi/

Como um agravante para esse quadro, a Egiptologia é uma disciplina particularmente exigente quanto ao domínio de bibliografia multilíngue. As obras essenciais, ou seja, o estado da arte e a discussão de todas as temáticas da disciplina obrigam o pesquisador a uma consulta de obras disponíveis em inglês, francês e alemão.

A produção egiptológica em língua portuguesa, além de mínima, é ainda pouco relevante e periférica em relação ao cenário internacional. A formação de um egiptólogo brasileiro exige, portanto, a pronta construção de ferramentas essenciais para o seu ensino na língua portuguesa, associado à aquisição paralela de idiomas estrangeiros. Deste modo, qualquer investigação sobre o Egito que não esteja instrumentalizada com idiomas estrangeiros ficaria circunscrita a referências muito limitadas.

É preciso reconhecer o esforço de acadêmicos brasileiros no ensino da língua egípcia que antecedem à publicação da primeira gramática para a escrita hieroglífica clássica em português (PEREIRA, 2014, 2016)[9]. Ciro Flamarion foi o primeiro acadêmico brasileiro a ensinar a língua egípcia nas universidades brasileiras. Criou e produziu um material de Egípcio Clássico e Neoegípcio e ministrou cursos a um grupo de alunos na Universidade Federal Fluminense, durante a década de 1990, muitos dos quais eram seus alunos de pós-graduação na época. O seu método, entretanto, ainda não foi publicado (SANTOS, 2012). Poucos profissionais replicaram o ensino do egípcio antigo, mesmo com a adição de materiais em língua estrangeira.

Os primeiros a estudarem com o Prof. Ciro foram os professores Marcos Caldas, Julio Gralha, Haydée Oliveira, Marcia Bezerra, Cláudio Prado de Mello e Nely Feitosa, ainda na década de 1990[10]. Alguns alunos de Flamarion Cardoso deram continuidade ao ensino de língua egípcia no Rio de Janeiro na década de 2000, mas com público restrito. A Profa. Nely Feitosa organizou cursos na Faculdade São Bento (2010-2013), na Unilasalle (2010-2015) e na pós-graduação da UFRJ em 2018. Marcos Caldas promoveu cursos na UFF (2003-2006) e depois na UFRRJ, como parte das atividades dos seus grupos de pesquisa, e na UFRJ (2018). Entre 2005 e 2006, Moacir Elias Santos ministrou um curso de extensão pela UNIANDRADE, no Paraná. Em 2007, o curso de especialização em História Antiga e Medieval das faculdades Bagozzi, de Curitiba, incluíram um módulo de Língua Egípcia na grade, também ministrado por Santos. O curso posteriormente passou para o ITECNE (atualmente Faculdades Madalena Sofia) e continuou na grade até 2016, quando a carga horária foi incorporada a uma disciplina destinada ao Egito Antigo. Em 2012, o curso de língua egípcia foi oferecido pelo ITECNE, como curso de extensão.

Vale destacar iniciativas do Centro de Estudos Interdisciplinares da Antiguidade (CEIA-UFF) e do Grupo de Estudos Egiptológicos Maat, ambos no Rio de Janeiro, que promoveram

[9] Em Portugal, a língua egípcia era ensinada na década de 1990 na Universidade Nova de Lisboa pela Profa. Dra. Maria Helena Trindade Lopes e na Universidade de Lisboa pelo Prof. Dr. Manuel Araújo tendo como principal referência a gramática de Gardiner (2012 [1927]).

[10] Agradecemos a Nely Feitosa e Moacir Elias Santos por essas informações.

cursos introdutórios. A XII Jornada de Estudos do Oriente Antigo - O Egito Antigo no Terceiro Milênio, em Porto Alegre, realizou um curso intitulado *Introdução a Mais Bela Escrita do Mundo: Hieróglifos,* organizado pelos professores Margaret Bakos, Júlio Gralha e Moacir Elias Santos. Entre as décadas de 2010 e 2020, o Seshat (Laboratório de Egiptologia do Museu Nacional) também ministrou cursos de extensão com caráter introdutório. Cássio Duarte Araújo, Maurício Schneider e Antonio Brancaglion Jr. ofereceram também cursos de introdução em São Paulo.

O Neoegípcio foi ministrado por Cardoso em 2010 e Liliane Coelho ofereceu o curso para alunos que integravam o projeto de pesquisa sobre as cartas de escribas de Deir el Medina, em Porto Alegre, coordenado por Margarete Bakos. Em 2015 e 2019, empregado os materiais do curso oferecido em Curitiba, Coelho também ministrou um curso de extensão sobre o Egípcio Médio, pelo Núcleo de Estudos da Antiguidade da UERJ.

O esforço de ensinar a língua egípcia ficou concentrado na região sudeste do Brasil, dificultando o acesso para alunos de outros Estados. Até o início da década de 2000, a internet estava ainda sendo implementada no país e o acesso era limitado. Ao mesmo tempo, a dispersão dos grupos de pesquisa e as dificuldades da realidade brasileira na formação de profissionais da Egiptologia fizeram com que poucos alunos seguissem no estudo sistemático do egípcio antigo. Nesse sentido, apesar da relevância da obra de Pereira, o impacto da sua gramática imediatamente após o lançamento, ainda foi limitado no Brasil, o que pode ser justificado pelo reduzido capital humano habilitado a ensinar a língua egípcia clássica.

Mas se o cenário da Egiptologia no Brasil parece bastante desolador até então, é preciso dizer que transformações significativas estão ocorrendo. A primeira delas foi o resultado de concursos públicos e da ampliação de vagas e instituições no Ensino Superior durante a década de 2010 que promoveu a descentralização do eixo Rio-São Paulo (SANTOS, 2017, 2019). Essa expansão, que atingiu todas as regiões do país, criou novos centros de pesquisa dedicados ao Egito antigo e o número de alunos integrantes de programas de pós-graduação também se ampliou graças à expansão de financiamentos para a pesquisa. Um outro elemento foi também o processo – lento – de internacionalização do Brasil na Egiptologia, com o retorno ao país de alunos de pós-graduação tendo formação parcial ou integral no exterior, que aprenderam a língua egípcia como parte de sua formação acadêmica possibilitando parcerias internacionais.

O aumento do número de pesquisadores mudou pouco a realidade da nossa formação, à medida existem elementos estruturais que não foram significativamente transformados (bibliotecas, acesso a periódicos, linhas de financiamento, etc.), mas tem criado oportunidades de colaboração, facilitadas recentemente pelas plataformas digitais. Nesse sentido, a situação da pandemia da COVID-19 acelerou um processo de divulgação de pesquisas, eventos e publicações, em grande parte promovidas pelas redes sociais. Isso também favoreceu a aproximação de pesquisadores de diferentes regiões do país (e do mundo), incluindo também possibilidades de participação do público não-especialista consumir esses conteúdos. É nesse escopo que precisamos examinar os benefícios e limitações das plataformas digitais.

2 As humanidades digitais e novos desafios para o ensino

Na última década, o desenvolvimento das Humanidades Digitais tem aberto novas oportunidades para as mais variadas áreas de pesquisa. Como uma área acadêmica interdisciplinar que fornece metodologias específicas da área das tecnologias digitais para serem incorporadas na investigação nas Humanidades como um todo (HOCKEY, 2004), essa nova frente de trabalho foi particularmente relevante durante o ano de 2020 com a pandemia da COVID-19 em que bibliotecas e instituições de ensino ficaram fechadas em quase todo o mundo pela maior parte do ano acadêmico.

Os inúmeros benefícios do trabalho interdisciplinar e as possibilidades de colaboração não devem diminuir, contudo, a importância das discussões teórico-metodológicas que são próprios das ciências humanas. Ao mesmo tempo, é preciso ter cautela com as promessas de inclusão e acesso aos conteúdos, levando em conta a variedade das realidades sociais e as particularidades de cada contexto de produção do conhecimento (KLEIN , 2015; BROWN, 2020).

Em linhas gerais, as Humanidades Digitais respondem a desafios profissionais e sociais também fora da academia (digitalização de património cultural, novas formas de cultura e de literatura, edição digital, criação de bases de dados, ferramentas de georeferenciamento, modelos em 3D, ambientes virtuais, tratamento de imagens, novos desafios de preservação em museus, arquivos, etc.). Mais do que apresentar uma definição sobre o que são as Humanidades Digitais[11], nos interessa aqui apontar como alguns dos seus recursos podem favorecer a ampliação de colaborações e do ensino. Nessa linha, Cohen sugere uma definição mais ampla, das Humanidades Digitais como um estímulo para o desenvolvimento de todo o potencial de investigação, ensino, publicação e divulgação das humanidades, através da incorporação de ferramentas, fontes e métodos digitais (COHEN, 2011). Alves (2016, p. 96), por outro lado, não a reduz a uma disciplina, mas a entende como uma comunidade transdisciplinar que amplia a possibilidade de trocas ao mesmo tempo que impacta as estruturas tradicionais em torno das formas de interação e mesmo da formulação de projetos. Mournier (2010) as trata como uma disciplina transversal, portadora dos métodos, dos dispositivos e das perspetivas heurísticas ligadas próprias do universo digital, mas adaptadas à realidade das ciências humanas e sociais.

Se por um lado as Humanidades Digitais englobam o conjunto de pesquisas e experiências que visam facilitar a utilização dos recursos digitais no âmbito das ciências sociais e humanas, elas não se limitam a uma simples transferência do meio. Trata-se de problematizar também o processo de construção do conhecimento. No caso da História, por exemplo, as plataformas digitais re-contextualizam as fontes primárias, e esse processo precisa ser problematizado e discutido. Noiret (2015, p. 33) nos lembra também que o desenvolvimento de

[11] Não há uma definição única para o termo. Ver por exemplo: EADH – European Associaiton for Digital Humanities (https://eadh.org/education/digital-humanities-centres); UCLCDH – UCL Centre for Digital Humanities (https://www.ucl.ac.uk/digital-humanities/); https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities/; http://www.iea.usp.br/en/news/digital-humanities-and-interdisciplinary

uma relação estreita com as tecnologias pode modificar os próprios parâmetros da pesquisa. É preciso reforçar que as Humanidades Digitais devem ser pensadas como uma nova forma de solucionar os problemas da investigação[12] sem deixar de lado a complexidade as múltiplas modalidades de análise em profundidade, crítica e interpretação (EVANS e REES, 2012; MALERBA, 2017; MOERBECK e ROCHA, 2021, no prelo).

Segundo Guerreiro e Borbinha (2014, p. 73), as Humanidades Digitais são inegavelmente um campo fértil da investigação. Todavia, uma das suas fraquezas é a ausência de modelos genéricos de sistematização (arquivamento) e reutilização (partilha) da informação reunida. A ausência de um modelo único, contudo, pode abrir espaço para o desenvolvimento do campo, acomodando demandas específicas de cada disciplina e seus problemas teórico-metodológicos. Alves (2016, p. 103) defende que Humanidades Digitais não necessitam de um processo de institucionalização formal, ou seja, que se transformem numa disciplina independente. A difusão e o aperfeiçoamento dessas ferramentas seriam então inerentes à atualização das humanidades diante de uma nova realidade tecnológica. O autor conclui:

> Talvez o conceito de comunidade e o desenvolvimento das suas múltiplas formas de afirmação, como se viu – da prática de investigação, à interação com o mundo para além da academia, passando pela construção dinâmica do conhecimento ou pelos novos métodos de validação do mesmo – possam fazer mais pela difusão e afirmação da qualidade, valência e relevância da investigação em Humanidades com uma componente Digital, do que as tentativas, por vezes forçadas e pouco estruturadas, de formalizar a sua presença no meio académico. (ALVES, 2016, P.103).

Tal posição reflete o modelo defendido por Cordell (2015), no qual as Humanidades Digitais deveriam ser incorporadas naturalmente ao currículo de cada disciplina acadêmica.

O desenvolvimento da internet teve um impacto significativo na produção e circulação do conhecimento. Os meios tradicionais da produção de saberes – o livro e a academia –, foram subvertidos ao mesmo tempo que houve uma transformação profunda na audiência que consome estes conteúdos. Mais ainda, o novo público deu forma a novos hábitos desse consumo, ampliando as possibilidades de interação (NOIRET, 2015).

A fim de garantir e expandir o acesso, é preciso repensar as formas e o tempo de consumo: textos e vídeos curtos, linguagem acessível, organização e distribuição visual das informações passam a ser submetidas a uma nova velocidade de atenção e resposta. O dinamismo e imediatismo do ambiente virtual trouxeram impactos significativos para o ensino e é preciso enquadrar o debate para além da polarização que coloca a internet como 'boa' ou 'má.[13]

---

[12] Ferramentas de processamento de imagem, criação e curadoria de grandes bases de dados são exemplos comuns que museus e bibliotecas especializadas têm utilizado na preservação de fontes históricas. A criação de modelos em 3D através da fotogrametria e de softwares de GIS também auxiliam a arqueologia no entendimento da paisagem e das fontes materiais. Técnicas não invasivas para obtenção e processamento de imagens têm sido utilizadas em larga escala para o tratamento de restos humanos, o entendimento de técnicas de enterramento, mumificação, etc.

[13] Ver o caso da Wikipedia e projetos de educação associados a ela, como por exemplo, Marques (2013, 2019), Louvem e Marques (2013), Moerbeck e Rocha (2021, no prelo).

3 Ensinando a língua egípcia em ambiente virtual

O projeto de ensino do Egípcio Clássico examinado neste artigo está ainda em fase experimental, mas teve resultados importantes no seu primeiro módulo em 2020, os quais serão discutidos a seguir e que devem servir de base para as próximas etapas. Nosso objetivo com este projeto é, por um lado, a consolidação da gramática como ferramenta de trabalho nas universidades brasileiras, somando-se à disponibilização de uma antologia de fontes e glossário ao acesso público, digital e gratuito. De outro, tem como ambição a promoção de um ambiente mais colaborativo entre os egiptólogos do país, à medida que pesquisadores podem compartilhar das mesmas ferramentas, recursos e multiplicar o ensino de língua em seus departamentos, consolidando o conhecimento tradicional da Egiptologia com os debates desenvolvidos no Brasil.

A promoção do aprendizado da língua egípcia antiga em Português é uma tentativa de acelerar a curva de aprendizagem e de minimizar os obstáculos de se aprender uma língua antiga em um idioma estrangeiro – experiência que os dois autores deste texto enfrentaram[14]. No caso específico de Pereira, que aprendeu todos os estágios da língua egípcia através de professores e bibliografia alemã, o processo, por vezes frustrante, de acompanhamento das aulas ajudou-o a desenvolver metodologias de estudo e a escrever os conteúdos gramaticais em português para estudo próprio. Tais metodologias eram completamente empíricas e foram ajustadas segundo a conveniência de cada lição aprendida. Esse longo processo de aprendizagem se desdobrou em atividades de ensino em Portugal, onde o autor teve a oportunidade de aplicar sua própria metodologia de aprendizado e aperfeiçoá-la como método de ensino em sala de aula desde 2011[15]. A ausência completa de material didático em língua portuguesa forçou-o a preparar apostilas para o ensino da gramática egípcia. Esse material era constantemente revisto e, eventualmente, expandido. Como resultado do esforço, em fins de 2014 foi publicada a primeira edição da primeira gramática de língua egípcia em português (PEREIRA, 2014, 2016)[16].

O principal objetivo dessa obra era permitir ao estudante lusófono o aprendizado da gramática egípcia na sua própria língua materna. Embora ainda não exista um dicionário de Egípcio-Português, o vocabulário existente na obra soma cerca de 900 palavras traduzidas, em adição às tabelas de pronomes, preposições, numerais e calendários. Há ainda uma lista com mais de 700 hieróglifos egípcios descritos e analisados na secção de apêndices, disponíveis para consulta. A gramática foi concebida como um desmistificador da língua egípcia, desfazendo a ideia de que, para se entender os textos egípcios, é preciso algum tipo de iniciação ou genialidade.

[14] O ensino de grego e latim já conta com a experiência de profissionais que têm ministrado aulas no ambiente virtual para pequenos grupos ou individualmente. A oferta de materiais em língua portuguesa é maior e o debate sobre iniciativas de ensino também tem sido feito por especialistas. Sobre outros projetos de popularização de grego e latim, ver o Projeto Minimus; também Sumares (2014) e Santos (2020).

[15] R. G. G. Pereira ensina a gramática egípcia clássica para turmas de graduação e pós-graduação em História na Universidade Nova de Lisboa desde 2011.

[16] A obra foi revista, corrigida e expandida em 2016, passando o texto para o novo acordo ortográfico. Entretanto, a Universidade Federal Fluminense (Brasil) já na década de 1990 utilizava o material de autoria do Prof. Dr. Ciro Flamarion Cardoso (UFF), que ainda não foi publicada.

Com base nessa experiência, o curso de *Introdução ao Egípcio Clássico (Egípcio Médio)* foi pensado para o público brasileiro numa tentativa de facilitar o acesso ao aprendizado da língua e possibilitar sua multiplicação através do ambiente virtual. Das etapas envolvidas no seu planejamento, destacamos duas frentes. A primeira diz respeito ao aspecto operacional, com o apoio institucional e a escolha das plataformas digitais para nas aulas síncronas e assíncronas. O segundo elemento importante foi a definição do público-alvo, tendo em vista as demandas da área no contexto brasileiro.

Inicialmente foi feita uma sondagem com os professores especialistas em Egito antigo no Brasil a fim de se ter ideia da possível demanda interessada, principalmente dos laboratórios de pesquisa espalhados pelo país. Tínhamos em mente atender alunos inscritos em programas de pós-graduação (Mestrado e Doutorado) que precisavam do conhecimento da língua para desenvolver seus projetos de pesquisa e também os professores especialistas que poderiam multiplicar o ensino da língua em seus grupos.

Com uma estimativa inicial dos colegas entre 20 e 30 alunos no máximo, de todo o Brasil, planejamos o curso em duas etapas. A parte assíncrona consistia num vídeo com a aula expositiva, baseado no conteúdo da gramática de Pereira e seguindo as lições e divisões de conteúdo apresentada no método.[17] O segundo momento da aula seria síncrono, a fim de garantir a discussão dos conteúdos, a correção dos exercícios e a possibilidade de prática de leitura a partir de fontes primárias (estelas, inscrições monumentais, funerárias, etc.)[18]. Paralelamente, outros materiais de apoio seriam disponibilizados ao longo do curso.

Com o apoio da Universidade Federal de Santa Catarina (UFSC) e do GTHA/ANPUH foi possível credenciar o curso como um projeto de extensão, emitindo certificado aos alunos, o qual poderia ser incluído em suas atividades de produção acadêmica. Num outro nível, a divulgação do curso aproveitou da rede de comunicação do GTHA/ANPUH (lista de emails, redes sociais), que inclui todos os professores de História Antiga no Brasil e muitos alunos. A divulgação em larga escala foi importante, mas deixou de lado outros interessados.[19] Na tentativa de ampliar o acesso e não prejudicar outros pesquisadores, foi acordado que o material das aulas expositivas ficaria disponível no canal do GTHA/ANPUH no YouTube, uma plataforma digital de fácil acesso e gratuita.

---

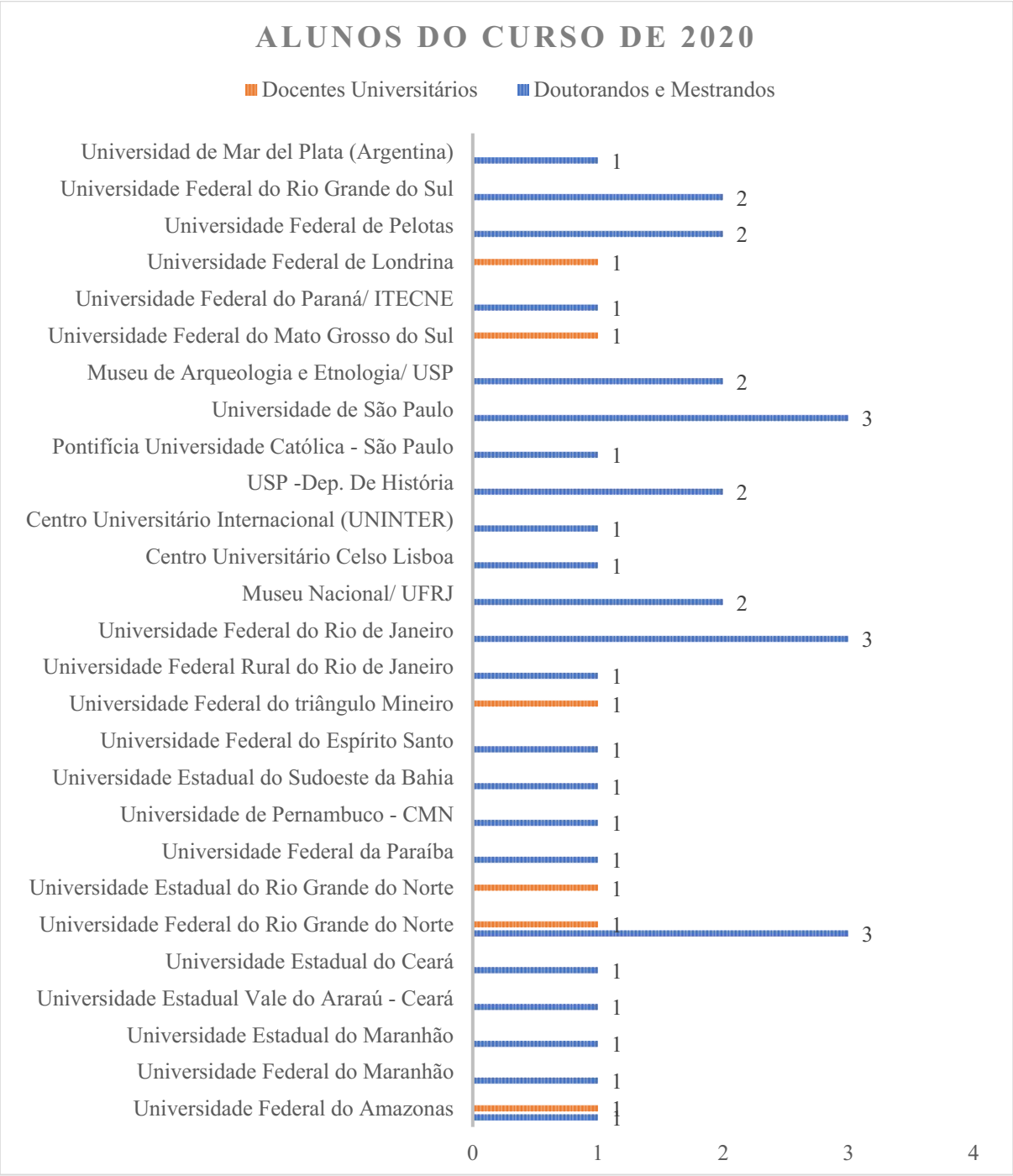
17 A aquisição da gramática era pré-requisito para o curso. A editora disponibilizou também a versão em e-pub. Por conta do contexto da pandemia em 2020, essa era uma opção mais barata e que não dependia da entrega dos correios, facilitando o acesso a todos os alunos.

18 Os encontros foram feitos pela plataforma *Google Meet.* O *link* para o encontro síncrono era enviado 15 minutos antes do início da aula. Os alunos recebiam as vídeo-aulas uma semana antes da aula síncrona e tinham este período para realizar os exercícios da gramática.

19 Apenas no primeiro dia de divulgação oficial, o projeto recebeu mais de 120 pedidos de inscrição. Com a alta procura, foi preciso fazer um processo de seleção que levasse em conta os projetos de pesquisa. Os critérios incluíam a afiliação a um programa de pós-graduação, projeto de pesquisa que envolvia o conhecimento de língua egípcia clássica e o estágio de desenvolvimento da pesquisa. Houve procura do público não-especialista e pesquisadores dedicados ao Egito greco-romano.

O curso (Fig.1) contou com 27 departamentos em 25 universidades (uma das quais, argentina, ver Gráfico 1). Foram 44 inscritos, divididos em duas turmas para as atividades síncronas. Concluíram o curso 32 alunos, respeitando o critério de presença das aulas síncronas (75%) e a realização da atividade de avaliação final, que consistia na transliteração e tradução de um texto de uma estela.

Figura 1. Instituições participantes do curso Introdução ao Egípcio Clássico (Egípcio Médio) – Formas Nominais.



Fonte: elaborado pelos autores.

Ainda que as vídeo-aulas sejam um elemento importante da aprendizagem, elas são apenas uma etapa do processo e devem servir como ponto de partida para o estudo individual. Sabemos da importância dos encontros síncronos e da necessidade de acompanhamento nos exercícios, principalmente para que os alunos adquiram segurança e familiaridade com a estrutura da língua a fim de lidar com a complexidade do processo de transliteração e tradução. As atividades síncronas tiveram como objetivo também o desenvolvimento de um repositório de fontes (estelas, portas-falsas, etc.) que constituem o principal material para o desenvolvimento das competências de leitura (transliteração e tradução).

Uma vez que tínhamos conhecimento dos projetos de pesquisa dos alunos, procuramos incorporar exemplos que pudessem ser úteis para seus respectivos projetos, discutindo também a importância do contexto e do suporte material das fontes egípcias, além da relação intrínseca entre texto e imagem. Os alunos tiveram, ao longo do curso, a possibilidade de explorar outros elementos das fontes, ainda que de forma inicial num curso introdutório, mas que permitiram ampliar o escopo das suas leituras e atividades de pesquisa[20]. Um elemento que consideramos fundamental foi a participação coletiva no repositório de referências: todos tinham acesso ao drive compartilhado e podiam fazer o *upload* de referências que fossem relevantes para o grupo.

Espera-se que este repositório seja ampliado no decorrer do módulo 2 (Formas Verbais) para, futuramente, constituir um banco de fontes e um glossário com acesso aberto em plataforma digital. Um dos elementos importantes para a concretização desse projeto é o fortalecimento da Egiptologia nas instituições de ensino superior brasileiras e de colaborações entre os especialistas lusófonos. O apoio do GTHA/ANPUH é também essencial a fim de otimizar os esforços e espaços colaborativos, como foi feito nessa primeira etapa.

Entendemos que o projeto foi um híbrido de atividades síncronas e assíncronas, como a maior parte do ensino durante o ano de 2020. Reconhecemos, nesse escopo, que um projeto de ensino de língua à distância requer outros recursos, tanto dos professores, como de ferramentas empregadas. As vídeo-aulas, aplicativos para memorização, dicionários *online* e o repositório para atividades e referências, sem incluir a gramática são *complementos* para a aprendizagem e devem ser vistos como ferramentas. Os encontros síncronos eram o momento crucial do curso onde se privilegiava a interação entre os colegas e destes com os professores. Contudo, a interação no ambiente virtual fica limitada muitas vezes ao tipo de suporte utilizado (computador ou celular, qualidade da câmera, microfones). Além disso, é preciso levar em conta um outro tipo de organização da comunicação (quem fala, quando fala, como sinalizar, como interromper), que é muito distinto quando feito presencialmente. Apesar da estabilidade da ferramenta, muitas vezes, alunos não conseguiram participar por problemas de conexão da internet e falta de energia elétrica.

[20] Muitos alunos nos procuraram individualmente com questões específicas de tradução das fontes que estavam trabalhando. Ainda que o curso não permitisse um conhecimento amplo e rápido do egípcio antigo, muitos se arriscaram mais com suas próprias traduções e a participação dos professores foi bastante pontual.

O potencial para o uso de plataformas e ferramentas digitais na Egiptologia já é conhecido e explorado no exterior. Dicionários, coleções de museus com acesso aberto, palestras e cursos em ambiente virtual são elementos importantes que facilitam colaborações e o processo de internacionalização. Mais ainda, é preciso considerar que muitos destes projetos pressupõem o acesso de conteúdos pelo grande público, de modo que a linguagem nestas plataformas não fica restrita ao público especialista. No caso dos conteúdos relacionados a textos, como dicionários e bases de dados, é preciso que o indivíduo tenha familiaridade com as fontes e conhecimento da língua e da escrita. Uma formação inicial que instrumentalize a leitura do texto egípcio não se restringe necessariamente ao ambiente universitário e pode ser realizada em outros modelos, como é o caso da *Egypt Exploration Society* (EES) em Londres, que há muitos anos vem oferecendo cursos de Egípcio Clássico para o grande público[21].

No mundo lusófono, os trabalhos das universidades têm pouco alcance com o público não-especialista. No entanto, o uso das redes sociais pelos pesquisadores tem trazido experiências de sucesso na divulgação de conteúdos egiptológicos, tanto para o público especialista como para os interessados. O Instagram e o Facebook, por exemplo, hospedam contas individuais e grupos de pesquisa que compartilham informações sobre o Egito antigo e a Egiptologia e que podem operar de forma também de forma combinada, ampliando o impacto de alcance[22].

Iniciativas de colaboração têm se mostrado frutíferas na nova geração de investigadores falantes do Português, no Brasil, em Portugal e com a progressiva aproximação entre os dois países. Se as dificuldades são compartilhadas na formação de profissionais da área e na escassez de recursos financeiros, é preciso que a internacionalização fortaleça a produção e a circulação dos conteúdos em língua portuguesa, trazendo benefícios a médio e longo prazo.

Considerações finais

O projeto aqui apresentado surgiu de uma oportunidade de fomentar um diálogo entre docentes e investigadores em diferentes estágios de formação na área de Egiptologia no Brasil. Nesse mérito, acreditamos que o curso *Introdução ao Egípcio Clássico (Egípcio Médio)* pode contribuir para repensarmos práticas de ensino e colaboração tanto entre pesquisadores

[21] O Reino Unido possui diversas sociedades que fazem um trabalho importante de divulgação do conhecimento egiptológico para o público interessado, convidando professores e palestrantes especialistas na área. O caso da EES é interessante, pois ela promove também pesquisa, escavações, conferências e um importante trabalho de com arquivos. Para informações sobre a história e os objetivos da fundação ver: https://www.ees.ac.uk/

[22] Destaco aqui as iniciativas dos laboratórios brasileiros MAAT (UFRN) e Seshat (MN/UFRJ) e das contas no Instagram dos pesquisadores Inês Torres (Harvard University; @umaegiptologaportuguesa), Rogério Sousa (Universidade de Lisboa; @egiptologiaflul). Daniela Martins (University of Liverpool; @waysofhorus) reúne e divulga informações de conferências, seminários e outras atividades promovidas por egiptólogos também em língua estrangeira, num trabalho semelhante feito pelo EEF (Electronic Egyptological Forum). O trabalho de Marcia Jamile no Brasil, promove o canal Arqueologia Egípcia que tem como principal foco o público não especialista (youtu.be/NxPYOMCAkKY).

brasileiros, mas também no ambiente lusófono. A apresentação da primeira fase do projeto deve servir para uma observação crítica do emprego de ferramentas e estratégias de ensino digitais voltadas para a capacitação de um quadro de docentes e investigadores em formação.

Consideramos a realidade brasileira, país que apresenta dimensões físicas continentais e a existência de dezenas de laboratórios e departamentos dedicados ao estudo do mundo antigo, onde a investigação de temáticas egiptológicas têm crescido nos últimos anos. A distância física entre as instituições, espalhadas pelo território brasileiro, já dificulta o contato presencial de seus integrantes, principalmente tendo em vista o custo para se viajar dentro do país, o valor das bolsas de Mestrado e Doutorado pagos pelas agências financiadoras e os recursos limitados para a realização de eventos e atividades acadêmicas, que em geral não incluem os alunos da pós-graduação. O quadro se agravou com a pandemia de COVID-19 em 2020, mas também abriu possibilidades para a comunidade acadêmica.

Entendemos que um projeto de ensino a distância requer recursos e treinamentos específicos, sobretudo com tecnologias digitais que promovam atividades de interação. Se por um lado os recursos para o ensino a distância ainda são escassos e problemáticos no Brasil, acreditamos que as plataformas digitais também podem favorecer a aprendizagem e a cooperação, principalmente para a Egiptologia, ainda em estado embrionário no Brasil.

Os recursos digitais utilizados no projeto (vídeo-aulas, repositório de referências e atividades, aplicativos de memorização, plataforma para encontros síncronos) cumpriram um papel importante, permitindo o ensino de língua egípcia no país em larga escala e a aproximação entre investigadores. Ao mesmo tempo, essas ferramentas não substituem a prática docente e a necessidade de investimento sistemático no estudo do egípcio antigo (memorização de vocabulário, formas nominais, entendimento da estrutura da língua, etc.), principalmente com a leitura das fontes. Estes recursos permitem o armazenamento e o fácil compartilhamento dos conteúdos, o que pode auxiliar colegas que se disponibilizem a ministrar essas aulas no futuro.

Agradecimentos

Agradecemos aos Professores Alex Degan (UFSC), Dominique Santos (FURB) e Fabio Morales (UFSC), coordenadores do GTHA/ANPUH pelo apoio incondicional neste projeto. Nely Feitosa (UFRRJ), pelas informações sobre os cursos e em especial ao Prof. Moacir Elias Santos (UNIANDRADE) pelo diálogo aberto e colaborativo no planejamento dos cursos de língua e por fazer a ponte com colegas egiptólogos no Brasil. Agradecemos aos alunos e colegas que participaram dessa iniciativa e têm nos ajudado no aperfeiçoamento deste projeto.

Referências

ALVES, D. "As Humanidades Digitais como uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português". *Ler História*, 69, 2016, p. 91–103. https://journals.openedition.org/lerhistoria/2496

BAKOS, M. M. (ed.). *Egiptomania: o Egito antigo no Brasil.* São Paulo: Paris Editorial, 2004.

BRANCAGLION, A. "As coleções egípcias no Brasil" In BAKOS, M. (org.), *Egiptomania: o Egito antigo no Brasil*. São Paulo: Paris Editorial, 2004, p. 31–41.

BROWN, K. *The Routledge Companion to Digital Humanities and Art History*. [s.l.]: Routledge, 2020.

CARVALHO, M. M.; FUNARI, P. P. A. "Os avanços da História Antiga no Brasil: algumas ponderações". *História,* v. 26, n. 1, 2007, p. 14–19.

COLLIER, M.; MANLEY, B. Introducción a los Jeroglíficos Egipcios. Ilustraciones de Richard Parkinson. Versión de José R. Pérez-Accino. Madrid: Alianza Editorial, 2000.

EVANS, L., REES, S. "An interpretation of digital humanities". In BERRY, D. M. (Ed.) *Understanding digital humanities*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2012, p. 21–41.

FRANCISCO, G. S. "O Lugar da História Antiga no Brasil". *Mare Nostrum,* 8, 2017, p. 30–61.

FUNARI, P. P. A. "Ancient Egypt in Brazil: A Theoretical Approach to Contemporary Uses of the Past". *Journal of the World Archaeological Congress,* 6, n. 1, 2010, p. 48–61.

GARDINER, A. H. *Egyptian grammar: being an introduction to the study of hieroglyphs,* Oxford: Griffith Institute, 2012 [1927].

GUARINELLO, N. "A Morphology of ancient History from a tropical, half-European Viewpoint" In FUNARI, P. P. A.; GARRAFONI, R. S.; LETALIEN, B. L. (org.), *New Perspectives on the Ancient World: Modern Perceptions, Ancient Representations*. Oxford: Archeopress, 2008, p.1–7.

GUERREIRO, D. M., BORBINHA, J. L. "Humanidades Digitais: Novos desafios e oportunidades" *Cadernos BAD - Informação. Sociedade. Cidadania*, n. 1, 2014, p. 63–78. https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1060

HOCKEY, S. "The History of Humanities Computing". In: SCHREIBMAN, S., SIEMENS, R., UNSWORTH, J. (Eds.) *Companion to Digital Humanities*. Oxford, Blackwell, 2004, p. 3–19.

KLEIN, J. T**.** *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field.* Ann Arbor: University of Michigan Press, 2015.

LOUVEM, O.; MARQUES, J.B. A Wikipédia como diálogo entre universidade e sociedade: uma experiência em extensão universitária. *Anais do XIX Workshop de Informática na Escola (WIE 2013)*: p. 70–79.

MALERBA, Jurandir . Os historiadores e seus públicos: desafios ao conhecimento histórico na era digital. *Revista Brasileira de História*. São Paulo, v. 37, nº 74, 2017, p. 135-154.

MARQUES, J.B. Representação e visibilidade do mundo antigo na Wikipédia: gargalos e soluções. *Revista do Museu de Arqueologia e Etnologia*, 32 , 2019, p. 2–17

MARQUES, J.B. Trabalhando com a História Romana na Wikipedia: uma experiência em conhecimento colaborativo na universidade. *Revista História Hoje,* 2, n. 3, 2013, p. 329-346.

MOURNIER, P. "Manifeste des Digital Humanities" in: Journal des Anthropologues 122 –123, 2010, p. 447–452. https://journals.openedition.org/jda/3652

MOERBECK, G.; ROCHA, T. Da Antiguidade ao mundo atual: as dimensões da História Antiga e os seus públicos. *In*: MELO, Rosilene; MENESES, Sônia; WANDERLEY, Sonia. *Coleção Ensino de História. Volume: História Pública e Ensino*. São Paulo: Letra & Voz, 2021 (no prelo).

NOIRET, S. "História Pública Digital". *Liinc em Revista*, v. 11, n.1, 2015, p. 28–51.

PEREIRA, R. G. G. *Gramática Fundamental de Egípcio Hieroglífico*. Lisboa: Chiado, 2014, 2016.

ROCHA, T. "Brazilian Egyptology. Reassessing colonialism and exploring limits" In NAVRATILOVA, H.; GERTZEN, T. L.; DODSON, A.; BEDNARSKI, A. (org.). *Towards a History of Egyptology. Proceedings of the Egyptological Section of the 8th ESHS Conference in London,* 2018. Münster: Zaphon, 2019, p. 127–146.

ROCHA, T. "Tropical Egypt: The Development of Egyptology in Brazil and its Future Challenges". In LANGER, C. (org.), *Global Egyptology: Negotiations in the Production of Knowledges on Ancient Egypt in Global Contexts*. London: Golden House Publications, 2017, p. 161–171.

ROCHA, T. "O sorriso da esfinge: reflexões sobre o ensino do Egito antigo no Brasil" In LEMOS, R. (org.) *O Egito Antigo. Novas contribuições brasileiras.* Rio de Janeiro: Multifoco, 2014, p. 279–299.

SANTOS, D. O. "Ensino de História Antiga no Brasil e o debate da BNCC". *Outros Tempos,* v.16, n. 28, 2019, p. 128–145.

SANTOS, D. O. "De tablet para tablet - novas ferramentas para a pesquisa e o ensino da História das culturas cuneiformes na era digital". *Revista Tempo e Argumento,* 6, n. 12, 2014, p. 212–241.

SANTOS, D. O; KOLV, G.; NAZÁRIO, J. J. "O Ensino e a Pesquisa em História Antiga no Brasil: reflexões a partir dos dados da Plataforma Lattes". *Mare Nostrum - Estudos sobre o Mediterrâneo Antigo,* 8, 2017, p.115–153.

SANTOS, D.O.; SANTOS, D. G. "O projeto Paideia: Ensinando grego antigo no município de Blumenau (SC)". *Nunt Antiquus*, Belo Horizonte, 16, n. 1, 2020 p. 193–218.

SANTOS, M. E. "Estelas, Hieróglifos e Ciro Flamarion Cardoso: Uma contribuição ao desenvolvimento da Egiptologia no Brasil". In: ARAÚJO, S.R.R; LIMA, A.C. (orgs.) Um combatente pela história: Professor Ciro Flamarion Cardoso. Rio de Janeiro: Vício de Leitura, 2012, p. 105–123.

SILVA, G. J. "Os avanços da História Antiga no Brasil". *Anais do XXVI Simpósio Nacional de História – ANPUH*, 2011, p. 1–31.

SILVA, S. C. "Aspectos do Ensino de História Antiga no Brasil: algumas observações". *Alétheia* 1, 2010, 145–55.

Links da Internet

COHEN, D. "Defining Digital Humanities, Briefly", 2011 – (acesso em 17/12/2020). http://dancohen.org/2011/03/09/defining-digital-humanities-briefly/

CORDELL, R. "How Not to Teach Digital Humanities", 2015 – (acesso em 17/12/2020). https://ryancordell.org/teaching/how-not-to-teach-digital-humanities/

EADH – European Associaiton for Digital Humanities - (acesso em 15/01/2020). https://eadh.org/education/digital-humanities-centres

EEF – The Egyptologist's Eletronic Forum - (acesso em 15/01/2020).

https://www.ees.ac.uk/

EES – Egyptian Exploration Society - (acesso em 15/01/2020).

https://www.ees.ac.uk/

EF English Proficiency Index – 2020 – (visitado em 19/12/2020). https://www.ef.com/wwen/epi/

Learning English in Brazil - British Council (São Paulo, 2014) – (acesso em 20/12/2020). https://www.britishcouncil.org.br/sites/default/files/learning_english_in_brazil.pdf

Instagram - (acesso em 20/01/2020).

Daniela Martins (University of Liverpool: @waysofhorus);

Inês Torres (Harvard University: @umaegiptologaportuguesa);

Rogério Sousa (Universidade de Lisboa: @egiptologiaflul).

Laboratório MAAT – UFRN - (acesso em 15/01/2020).

http://maatufrnhome.blog/

Laboratório SESHAT – MN-UFRJ - (acesso em 15/01/2020).

https://seshat.museunacional.ufrj.br

PEREIRA, R.G.; ROCHA, T.; DEGAN, A.; SANTOS, D.; MORALES, F. Como se faz um Egiptólogo? Uma conversa com Thais Rocha da Silva e Ronaldo G. Gurgel Pereira – (acesso em 20/12/2020).

https://www.youtube.com/watch?v=khm1yrvGhvA

PEREIRA, R.G.; ROCHA, T. Introdução ao Egípcio Clássico (Egípcio Médio) – Formas Nominais.

https://www.youtube.com/playlist?list=PLI8rGh6UbR_vOBaIrALDQwiSHSgQ2TleQ

Projeto Minimus, ano II: *O grego e o latim no Ensino Fundamental Projeto de Cultura e Extensão da PrCEU da USP*. Programa de Pós Graduação em Letras Clássicas (FFLCH-USP), 2019. http://ppglc.fflch.usp.br/node/390

SUMARES, G. "Nova Vida para as Línguas Mortas". *Revista Cultura e Extensão USP*. Setembro, v. 11, Suplemento, 2014, p. 12-16 - (acesso em 15/01/2020).

http://www.calameo.com/read/001358971215118480e98

UCLCDH – UCL Centre for Digital Humanities - (acesso em 15/01/2020).

https://www.ucl.ac.uk/digital-humanities/

https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities/
http://www.iea.usp.br/en/news/digital-humanities-and-interdisciplinary

YouTube

Marcia Jamile "Arqueologia Egípcia" – (acesso em 20/01/2020).

https://www.youtube.com/channel/UCGv1ImAroULYoQTCt1RpzxA

Recebido: 02/02/2021.

Aprovado: 16/03/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 83-102, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p83-102

Article / Artigo

Building Digital Humanities on the Linguistic Background: Methodological Basis for Digital Humanities Education in Undergraduate and Graduate Programs

Construir Humanidades digitais num contexto linguístico: bases metodológicas para o ensino de Humanidades Digitais no 1º e 2º ciclo de ensino universitário

Lukáš Zámečník[*]
lukas.zamecnik@upol.cz
https://orcid.org/0000-0001-8098-4238

Ľudmila Lacková[**]
ludmila.lackova@upol.cz
https://orcid.org/0000-0001-9852-4280

Abstract

The transformation of society towards digitalization and automatization cannot be ignored by the higher education system. While this has been naturally reflected by the education system regarding technical sciences, humanities are still struggling to catch up with the latest trends in the digitalization of society. The field of Digital Humanities (DH) is very young and lacks a solid methodological basis or ontological principles. This paper aims at proposing some philosophical and methodological grounding for the field of DH and its practical applications in the higher education system. We describe two case studies of the creation of new study programs at the Palacký University Olomouc, Czech Republic.

Keywords: Digital Humanities; Education; Methodology; Philosophy of Science; Quantitative Linguistics; Study Programs.

Resumo

A transformação da sociedade em direção à digitalização e automação não pode ser ignorada no sistema educacional superior. Embora o processo tenha vindo a ser naturalmente refletido pelo sistema educacional de ciências técnicas, as Humanidades continuam a enfrentar dificuldades para se manterem atualizadas com as últimas tendências da

[*] Department of General Linguistics, Palacký University, Olomouc, Czech Republic
[**] Department of General Linguistics, Palacký University, Olomouc, Czech Republic.

digitalização da sociedade. A área das Humanidades Digitais (HD) é muito jovem e carece de uma base metodológica sólida ou de princípios ontológicos. Este artigo tem como objetivo propor alguns fundamentos filosóficos e metodológicos para a disciplina das HD, e respectivas aplicações práticas no sistema de Ensino superior. Descrevemos dois estudos de caso sobre a criação de novos programas de programas de estudo na Universidade Palacký em Olomouc, República Tcheca.

Palavras-chave: *Humanidades digitais; Educação; Metodologia; Filosofia da Ciência; Linguística Quantitativa; Programas de Estudo.*

Introduction

The history of the field of DH has its origins in the practical demands of philologies and the humanities based on them (LEE, 2018). It is based on the tradition of digital word processing. The symbolic birth of Digital Humanities is considered to be the decade-long digitization of the work of Thomas Aquinas by the priest Roberto Busa (started in 1946). In the case of Busa's project, we cannot talk yet about the digitalization in the sense of the today's meaning of the word, but it was a project of lemmatization of the data with a semi-automatic process (see BUSA, 1980). Digital Humanities can be classified alongside Digital Social Science in the broader category of Data Science. All these post-disciplines are conceptually based on the theory of dynamical systems (KELLERT, 2008). This theory (built since the 1960s) makes it possible to explain the behavior of complex systems regardless of their ontology, by capturing common features or isomorphic (or analogous) structures across similar systems.

Some mathematical models are created using statistical methods, with the goal to find some common and stabile characteristics of data. The universality of the found statistical distributions of data (in diverse ontologies) led to the definition of some principles of the theory of dynamical systems (especially in the context of the so-called scale-free networks, see CALDARELLI, 2007). Until recently, the main limitation of the development of the theory of dynamical systems for the needs of the systems described in the social sciences and humanities was the lack of data. At present, however, computer systems based on machine and deep learning methods (namely artificial intelligence) provide tools for extracting a growing amount of information from the Internet (social networks, etc., so-called Internet Artificial Intelligence), from the digital archives of government sections, from hierarchical lists of corporate data repositories (so-called Corporate Artificial Intelligence), and from the emerging data packages of the emerging Internet of Things (so-called OMO, online-merge-offline, so-called Sensory Perception Artificial Intelligence). In the near future, data collected by autonomous AIs (smart homes, autonomous vehicles, etc.) will also be added (see mainly the chapter "The Four Waves of AI", in: LEE, 2018, p. 104–139).

Digital Humanities is one of several interesting outcomes of the history of the humanities, starting with the turn to the Geisteswissenschaften and ending with Foucault's

efforts to build "human sciences" (FOUCAULT, 1966). In this sense, Digital Humanities are the culmination of the ambitions of the Archaeology of Knowledge (FOUCAULT, 1969). Where historical (and then social) causality exists for Foucault, but is not traceable and is therefore useless as a concept, there is a massive pattern recognition ability in Digital Humanities using AI (finding correlations to an extent that is inaccessible to humans) which overcomes the limitations associated with the search for simple causal relationships. The Archaeology of Knowledge is based on the AI's ability to extract data and with the help of heuristics, which often escape explanations, provide their interpretation, recognize a stable pattern of behavior (consumers), actions (social agents), or decision-making (judges, teachers, etc.).

1 Integrating DH in education

The still growing tendency towards creating new study programs in DH is motivated by the changes in society during the Fourth Industrial Revolution, especially by the expected changes in the labor market (the DH Lab in University Nova of Lisbon, a two-year master's program Digital Humanities and Digital Knowledge at University of Bologna and many others). Expressed by means of the theory of dynamical systems, there is a real risk that the self-regulatory mechanism of the free market will eventually cease to function, because cheap labor will cease to be a competitive advantage. According to some analysts, AI technologies will lead to further growth of corporate structures of the global economy. This is not an ideological statement, but a statement of a probable change in system dynamics. A change in several key regulatory parameters will lead to a change in the attractor of the global economic system. In order to prepare our society for these changes, we should accelerate the adoption of technologies. According to the strategical document Digital Europe Programme, EU member states should try to accelerate the adoption and best use of digital technologies, including the latest digital capacities, across the economy and society. According to this strategic document, "all Member States can identify, analyse and adapt to digital trends, establish the needs and priorities of the public and private sectors, share best practice, and contribute to common specifications and standards"(Digital Europe Programme Draft orientation, p. 29).

Ross (2016) and Lee (2018) show that in the developed world, not only manual-based sectors will be affected (this will be the area in the developing world where the advantage of cheap labor will have dramatically reduced), but also some sub-occupations in the services sector, which are also trained through the faculties of philosophy, will probably be significantly automated (such as some areas of media, interpreting, translation, etc.). In addition, automation will affect the industry of mental activities (e.g., leading positions in company offices, customer services, tax consulting, etc.). As an example, we can mention the eGrants from the European Commission, where the whole process of project submissions, evaluations, etc. has been automatized to a great extent, mostly during the Covid pandemics period (see the document Strategic Plan 2020–2024 - Research and Innovation, p. 43).

AI technologies are undeniably replacing some of the historically existing work positions at the labor market. Nevertheless, at the same time, new positions requiring a human interpretive approach are arising. As a consequence of the great progress in the developing technologies – and as a paradox – there is new room for a person. Not all patterns automatically recognized by AI are equally relevant, and it is the human mind that should be able to assess their relevance and usefulness. As a matter of fact, the field of AI is probably at the beginning of the third-wave AI, but the vision of an artificial agent able to make decisions in the way that a human being (genuine intelligence) does is far from close to be achieved (SMITH, 2019). In the current state of DH, the great optimization ability of AI at the level of quantity is connected with the optimization ability of a person at the level of quality.

This is the main motivation for the need of investments in the new study programs of DH. According to the strategic document of the European Commission Strategic Plan 2020–2024 – Education, Youth, Sport, and Culture,

> technology and the future of work, digitalisation of society and learning, or the transition to a circular economy necessitate that education and training systems across Europe can deliver the knowledge and skills, including digital skills and sustainable education that people need to participate fully in society (EUROPEAN COMMISION, 2020, p. 16-17).

Scholars are still more aware of the fact that the more technology advances, the more we need to stick to what makes us different from machines. This is why not only the education in humanities and social sciences is being radically modified in the form of the common name of DH but also the education in technical sciences and engineering started to be questioned and modified according to the new conditions of the Fourth Industrial Revolution. Some institutions are trying models of a revolutionary change of the education in engineering. For example, Texas Tech University in Lubbock launched a project (it is called DREAM: The Developing Reflective Engineers through Artful Methods) which already shows positive results in the education of engineering programs. The core idea of the change in the education of technical programs is that in order to be a more effective engineer, creative thinking (*Artful Methods*) is an indispensable predisposition. Insertion of artistic and creativity-based courses (such as creative writing) in the curricula of engineering programs proves positive outcomes of the students in the field (CAMPBELL *et al.,* 2020).

Nonetheless, there is also a counterpart: not only programs in the humanities are being digitalized and more *technical*, but also technical study programs are being *humanized*. This fact demonstrates the need of investment of energy and funds into quality education at the intersection of both directions. In this way, we will be ready, and the next generation will be prepared for the symbiosis with AI agents in everyday life. But not only that, the next generation will be prepared to face the changes of the labor market. The transformation of the labor market will largely affect graduates of humanities in general and mostly the graduates of philological disciplines – the small philologies as well as the large ones. Therefore, the overall integration of digitalization in the education of humanities is needed more then ever before.

2 Linguistic Digital Humanities

As was already mentioned in the introduction, the field DH originated in philologies as a necessity to preserve old manuscripts in the digital form (with the first digitized opera being the work by Thomas Aquinas).

Since the digitization of Thomas Aquinas' texts, the development of DH has continued in the direction of digitizing other important philological manuscripts or prints. Besides the transcription of old texts with aid of modern technological tools, the discipline of DH has nothing more in common with philology itself. It can be said that DH in the most general meaning is, strictly speaking, nothing but methodology – the researcher has digitized data, sets of computational methods, statistical tools, etc. and perceives them as tools that can be used in traditional disciplines (e.g., digital competences in medieval studies, in film studies, in gender studies, etc.). It can be said that DH conceived in this way are only a means of modeling data (GIERE, 2006, p. 68-69). Only with a linguistic background can we speak of DH as a new own-standing discipline that puts these data models into a theoretical framework and consequently allows their interpretation. We believe that the above-mentioned understanding of DH as pure methodology can be generalized, and it can be said that some areas of data science, despite their close affiliation with the natural sciences, remain mere clusters of methods. And the extent to which they can fulfill a role analogous to linguistics in another discipline (computational science) determines their future. In the humanities, we can expect greater integration through DH, and perhaps even the completion of the Foucault's project, as indicated above.

Merging the digital tools with linguistic theory, the discipline of DH can get closer to the original philological direction. Despite all the differences between the two authors, both Foucault and Derrida (1976) spoke about the central role of language (or text) in the scheduling of the humanities (or less categorically: *in capturing human destiny*). And even in the critique of the overly descriptive nature of their approaches, of their constructivism and discontinuism, of the relativism they have left room for, the central role of language still remains crucial in their theories. However, we are talking about language not only philosophically assumed, but more importantly, a language precisely defined and described by linguistics.[1]

The concepts of artificial language and natural language are central for DH – methodologically (artificial programming languages in methods), ontologically (DH are in most cases dealing with textualizable objects[2]) and axiologically through linguistic interpretation in a semiotic way.[3] Methodologically, because computational methods that allow data processing are artificial languages whose specific algorithmic set of rules represents the grammar of the code. Ontologically, because the basic elements of the DH ontology are objects in their

[1] Both *Archeology of Knowledge* (FOUCAULT, 1966), and *Grammatology* (DERRIDA, 1976) refuse strict formal aspect of structuralism.

[2] We have to keep in mind ontological commitments of linguistics, see Quine (1953).

[3] Following Larry Laudan in his view of three-fold structure of paradigm (see LAUDAN, 1998).

textualized form. The choice of these objects is pragmatic as only textualized objects are accessible through the methods of artificial grammar processing. Axiologically, since the goal of DH is to map interpersonal interactions in all their complexity. These interactions are understood as the process of establishing, keeping, and transforming characters in the communication interchange.

Central to DH is also the concept of speech. However, not in the traditional structuralistic way of the term *speech* as in contradiction to *language* but in a new form that leaves room for the study of speech diachronicity in its continuous transformation. The original structuralist preference for synchrony over diachrony, as well as the subsequent poststructuralist rejection of the system, is overcome in the equal position of synchronous and diachronic ways of studying phenomena in DH. Following Veyen's Foucault[4], we can say that DH makes it possible to illustrate how the discourse is entrained by the dispositive. It will be possible to model this entrainment for individual disciplines in humanities that work with the historical dimension, similarly to the mood entrainment on social networks. In this sense, DH represents a natural continuation of corpus linguistics and text linguistics going in the direction of the developing of technologies applied to the research of authentic texts. The traditional methods of text linguistics (KOŘENSKÝ, 2003) are being extended and accompanied by various tools for automatic text processing, using corpora both of spoken and written speech.

We comprehend three core elements of Linguistic Digital Humanities (LDH): the basis is the segmentation of textualized objects accompanied by indispensable qualitative linguistic concepts and finally analyzed by digital tools for the analysis which lead to numerous possible applications. We comment accordingly on the three core elements in the following part of this section.

2.1 Segmentation of textualized objects

The exaggerated postmodern statement that "everything is text" has taken on a new message in the context of DH – we must treat textualizable objects as constructs composed of constituents – linguistic planes consequently set the frameworks for segmentations of a given type (morphological plane, syntactical plane, etc.). The determination of units – the basic components of segmentation and their constituents – is governed by some type of compositional principle[5] in which quantitatively traceable universal principles of scale-free networks can manifest themselves.[6]

[4] See Veyne (2010, p. 92-110). See Conspiracy Pedagogies: QAnon, Social Media, and the Teaching of Far-Right Extremism. In: ihr.asu.edu/seed-grants/conspiracy-pedagogies

[5] For the relation between the register hypothesis and the concept of power law see Köhler (2012, p. 84-92); for the principle of compositeness see Hřebíček (2003).

[6] Caldarelli (2007); Ferrer-i-Cancho; Solé (2001).

The linguistic areas that we consider to be very important for DH include computational and quantitative linguistics – universal statistical tendencies in text analysis from Herdan (1966) to Altmann (1978) and Köhler (1986); and corpus linguistics (see JENSEN, 2014) – also works on extracting semantic relations from language corpora, for instance extracting semantic relations from Portuguese corpora (AMARO, 2014).

2.2 Qualitative concepts

As already mentioned, if DH is not to remain a mere group of methods, it is necessary that theoretical tools are involved – concepts, hypothetical principles, but also the whole deductive structure of theories. Theory building in DH inevitably entails the need to work with qualitative concepts of linguistics.[7] The neglecting of qualitative theoretical tools and over-reliance on machinery of technology (statistics in particular) can also cause misleading results. An essential example is the importance of lemmatization – experiments with non-lemmatized data may be valid in some cases (e.g., when looking for a distribution function that expresses the relationship between the length and frequency of a lexical unit), but in others it leads to completely invalid conclusions (e.g., if we want to express the relationship between many meanings of a lexical unit and the length of the lexical unit[8]). These facts have long been known in linguistics (see KÖHLER, 1986) but are sometimes ignored by some current analyses within data science.

2.3 Tools for analysis, comparative tools and aplicability

Merging elements (1) and (2) we can obtain useful tools for analysis in many disciplines. One of the possible applications results in the creation of the new generation of vocabularies or glossaries. Terminology in the traditional sense (CABRÉ, 1999) of the discipline is being developed thanks to the possibility of big data analysis. The digitalization of the discipline of Terminology facilitates creation of electronical vocabularies, dictionaries, or glossaries for specific disciplines or for specific research projects. Besides the creation of new glossaries, the discipline can be also comprehended in the sense of digital editions of already existing vocabularies or dictionaries (SALGADO; COSTA, 2020) or creation of new digital tools for lexicographers (SALGADO; COSTA, 2019). Not only the access to a big amount of data helps the discipine of Terminology to become more „digital", but also the digital outreach of the discipline started to have an important impact on the society. We can mention the ongoing project Glossário Colaborativo COVID-19 by colleagues from the University Nova of Lisbon

[7] Meyer (2002), Grzybek (2006).

[8] There is a long discussion about the appropriate variants of measuring the lenght of linguistic unit, for the new approach see Benešová; Faltýnek; Zámečník (2015). The realtion between the lenght of lexical unit and the number of its meanings (including the role of lemmatization) is explicated in Köhler (2005, p. 767-770).

developed as an orienting tool – a terminological glossary for public use regarding the pandemics. The aim of the project is to help increase the public knowledge about the disease, virus, and pandemics in general. Wordnets (for Portuguese see for instance AMARO, 2006; AMARO *et al.,* 2013) are other tools for working with lexemes or terms, depending on the area in question. Another possible application sphere is automatic sentiment detection, potentially usable in robotics, chatbots, or similar spheres of AI (LESCH, 2015). One more possible example of applicability of Linguitic Digital Humanities is Forensic Linguistics (see FALTÝNEK; MATLACH, forthcomming). In the field of forensic linguistics, the most common application of the digital tools is the automatic recognition of authorship. This application field has begun to grow recently in the Czech Republic, and it became a research topic of many projects in cooperation with students at Palacký University in Olomouc (JANEČKOVÁ *et al*., 2021). Distant Reading represents another possible application, where stylistic, periodization, genre and other categories are studeied through Corpus Linguistics analysis of big data.

Conclusion

DH point a new direction not only for humanities, but also for social sciences, and the consequences of this new direction have started to be incorporated in the educational system (see the following section Case Studies). Since the field of DH is rather young and is still defining its status within the scientific and academic environment, there is no unification so far regarding methodology or the very philosophy of the filed. This fact might be considered as an advantage for everyone who is interested in the field: every scholar contributing to the field of DH can help in defining and shaping the discipline with his/her proper approach. We also take this opportunity and propose a unique, linguistic, understanding of DH. In our understanding, DH is not a mere methodology for already existing disciplines. DH represents a whole specific approach to every textualizable research object. This approach is composed of three main components (segmentation as a quantitative view, qualitative linguistic analysis, and digital tools for specific applications).

We tried to present here the educational effort or even goal of DH as the main and most important part of the initiative across *human sciences* and *digital knowledge* integrated in the project of DH. However, the general effort must be implemented in some concrete activities which can transform the boundary field of humanities as well as digital and data science. Concrete projects should be prepared, like our new study programs in Linguistics and DH, which will incorporate the *linguistic basics* into the educational perspective of DH.

Case studies

At the Department of General Linguistics of Palacký Unviersity Olomouc, we incorporate most of these linguistic areas into teaching strategies and build new curricula in which the DH affinities to linguistic subdisciplines have an important role to play. In the sense of the aforementioned preliminaries of *Linguistic Digital Humanities*, new study programs have been created under the name *Linguistics and Digital Humanities*. Two were accredited in the last two years: a three-year bachelor program and a four-year doctoral program. In the next section, we will shortly present both of the newly created study programs at the Palacký University Olomouc.

Case study 1: Bachelor program Linguistics and Digital Humanities

The bachelor study program Linguistics and Digital Humanities got accreditation in 2020 and we expect the first round of applications in the year 2021. It is a pilot case of a bachelor program in DH in the Czech Republic.[9] The program is largely focused on future applicants to the faculties of arts and humanities in general. It will provide them not only with basic theoretical knowledge, but above all with rich equipment of digital tools and quantitative and qualitative methods that will allow the graduates to regain a competitive advantage. Mastering the methods of work in a digital environment, the ability of software programming and orientation in databases will allow graduates of the study program to successfully enter the transformed labor market (as described above). For example, a translator who is able to integrate software tools into his work and at the same time further develop the very software will still have an advantage over mere automatic translators available to consumers as well as over human translators working with traditional CAT tools.

The latter role of DHin the case of research shows that graduates of the study program *Linguistics and Digital Humanities* can also move towards further study, which will focus on the professional profiling of students with research ambitions. Further research career of the graduates of this program is possible both under the auspices of social sciences and humanities, as well as within the theories and methods of DH themselves.

[9] The study program Linguistics and Digital Humanities fulfills the intention of the Czech Republic in the Olomouc Region to support the employment of university graduates in the areas affected by the Fourth Industrial Revolution. The preparation of the study program was supported by a grant from the European Union from the European Regional Development Fund, in the INTERREG program. The project is entitled "Digital Humanities for the Future", CZ.11.3.119 / 0,0 / 0,0 / 18_031 / 0002217. The project is implemented in cooperation with the University of Wroclaw, its Institute of Information Studies and Librarianship. Polish partners will also implement a part of the teaching in application courses.

The study program *Linguistics and Digital Humanities* is based on three pillars: (1) theory of Digital Humanities, (2) methods available for Digital Humanities, and (3) specific applications of Digital Humanities mainly from the field of general linguistics and individual philologies.

The theory pillar (1) is the subject of basic compulsory courses of the theoretical basis. These courses are: Theory of Humanities 1 and 2, Semiotics, Introduction to General Linguistics, The Past of Database Systems and The Present of Database Systems, Critical Discourse Analysis, and Forensic Linguistics.

Being a post-discipline, DH share most of their theoretical foundations with other disciplines and subdisciplines – their own autonomy is achieved through their specific interconnection / amalgamation. Theory of the Humanities 1 draws mostly from philosophy, both from the specific philosophies of the first half of the 20th century and from the philosophy of science (but also historiography, sociology of science, etc.). The connections between Humanities Theory 1 and 2 represent Michel Foucault's theoretical concepts. The Theory of Humanities 2 also draws from philosophy, namely from the critique of postmodernism and naturalism, and from the theory of dynamical systems, whose theory has become the basis for the creation of a specific theory of Digital Humanities. Introduction to General Linguistics and Semiotics represent the linguistic and general semiotic theoretical foundations of Digital Humanities. Basic theories of general linguistics, from structuralism and generativism to psycholinguistics, systems-theoretical linguistics, or cognitive linguistics, are a necessary source of concepts and hypotheses on which to understand the concepts of DH. Semiotics as a general theory of meaning is the most important qualitative contribution to the interpretation of data that make the methods of Digital Humanities more effective.

The courses Past and Present of Database Systems present various ways of organizing knowledge in the past and present. One of the key theoretical findings of DH is the idea of conceptual framing and organization of knowledge. Here, DH draw mainly from terminology and information science, but also from history and historiography. Critical Discourse Analysis and Forensic Linguistics will present two different conceptualizations of ways of analyzing language data (specific discourse and specific language corpora).

The pillar (2) is the subject of courses: Data processing in DH 1 and 2 and four courses of the profiling basis – Formal language processing 1 and 2, Quantitative language processing 1 and 2.

The Data Processing courses in DH 1 and 2 will introduce students to both the basic and advanced knowledge of statistical analysis, to the issue of data visualization, and also to interpretation of such analyses. Apart from that, the students will be acquainted with natural language processing (NLP), image processing, exploitation of social networks, etc.

The courses Formal Language Processing 1 & 2 and Quantitative Language Processing 1 & 2 will provide students with the knowledge of the basics of formal language analysis and

of the creation of text-processing algorithms. The outline of the whole series is designed in such a way that the students will deal with increasingly complex tasks: from formal word processing to algorithmization. The main objective for them is to efficiently exploit prefabricated applications to solve their tasks. The series of courses is completed by the elaboration of a year project in which students apply a set of acquired skills and knowledge. The natural continuation of this project will be the project of the bachelor's thesis itself.

The pillar (3) is the subject of courses: Applied Semiotics, Text Attribution, Creative Visuality, Natural Language Processing, Mathematical Modeling 1–3, and other optional courses in the profiling basis.

The courses, which represent a certain section of possible applications of Digital Humanities, will be continuously supplemented in connection with the development of the discipline, the expansion of the study program portfolio, and the expansion and transformation of the team of lecturers researchers, and interns. Due to the existence of the doctoral study program *Linguistics and Digital Humanities*, the composition of applied courses will also be based on the teaching activity of doctoral students of the mentioned program. Last but not least, the list of applied courses will be based on the offer of our Polish colleagues from the Institute of Information Studies and Librarianship, who will be involved in teaching.

Case study 2: PhD program in Linguistics and Digital Humanities

The DH doctoral program at the Department of General Linguistics of Palacký Unviersity Olomouc reflects a specific approach to this scientific discipline, which is based on methods of quantitative linguistics. Thanks to this methodology, students are able to examine texts on the basis of a wide range of qualitative and quantitative properties and with the requirement to process big data. The students become familiar with various software programs for text analysis, one of them being the QUITA software created at the Department of General Linguistics of Palacký University Olomouc (see KUBÁT; MATLACH; ČECH, 2014). QUITA evaluates a wide range of quantitative text properties, such as entropy, type-token ratio, average word length, etc., allows work with various text transformations, such as n-grams, hapax legomena, bag-of-words model, reduction, randomization, etc., provides visual representation, and serves data mining. At the same time, the fact that it can be used for the analysis of data from genetic banks demonstrates the possibility of extending linguistic methods into new areas within DH. QUITA software is now being used by a wide range of quantitative linguists, and its use has resulted in more than 60 professional studies that have moved the research in quantitative linguistic significantly forward (LIU-LINAG, 2017; POPESCU *et al.,* 2017; CHEN; LIU, 2018; GLOGAROVÁ–KUBÁT, 2020, and many others). It has also been used in dozens of diploma theses (latest ones: ZÁVODNÍ, 2020; VARMUŽOVÁ, 2020).

The program is based on three main research orientations (profiling lines). Each student decides for a specific orientation in relation to the topic of the dissertation: The three orientations are:

a. linguistic description for analysis of digitalized text,

b. analysis of digitized text for use by humanities (philology, history, etc.),

c. linguistic analysis of genetic text.

All of these orientations involve working with data mining methods.

The first of the profiling lines of study is focused on the use of linguistic description of language, text and its properties used in combination with methods of data mining and natural language processing (automatic text attribution for example). The language features that enter the analysis include, among others, grammatical and lexical categories. Such an analysis might further yield tools for a wide variety of linguistic disciplines (text theory, stylistics, pragmatics, etc.). The aim is to associate methods that lack a uniform *tertium comparationis* – are based on a different view of language and text – but as a union can very pregnantly express the specifics of individual texts either under academic research or under assessment in the application sphere. The courses aiming at the application sphere are mostly Forensic Linguistics and Linguistic Applications – those are developed and taught in collaboration with Institute of Formal and Applied Linguistics at Charles University in Prague (taught by doctor Kateřina Lesch). The courses of Programming and Corpus Linguistics (taught by prof. Amaro from University Nova of Lisbon) are also crucial for this line of study.

The second profiling line of study is focused on the use of linguistic analysis of the text within the research of the humanities. The aim is to implement the methods of data mining based on linguistic analysis in the research in their individual disciplines. At the same time, this approach assumes that it can be supplemented by a description of the research topic from the given humanities discipline – it will associate the methodology of the given discipline with an integrated linguistic description and use them together in data mining tools. Semiotics will play a protective role here, which will enable the description of the text, cultural artifacts, social phenomena, etc. to be viewed in a uniform characteristic framework – which a successful analysis of the studied phenomena presupposes. In addition to Semiotics, Terminology represents another unifying approach to the wide range of disciplines. Terminology in the above-mentioned sense (see the previous section) is an important part of DH in the second profiling line of study in that it is a discipline with wide-range applicability across the humanities – but not only. The course of *Terminology and organization of knowledge* is taught by prof. Rute Costa from the University Nova of Lisbon.

The third line of study is focused on the transfer of linguistic methods to the analysis of genetic text. The initial premise is again a unified semiotic framework – the concept of the genetic code, the structure of the genetic text, sign, and its function. The aim is to use linguistic

analysis of text together with data mining methods, in this case for the analysis of biopolymer chains – DNA / RNA and proteins. The linguistic methods included in the analysis of the genetic text will be presented in the study as corpus approaches and approaches verifying the manifestations of linguistic laws and quantitative metrics of the text in the genetic strings. Attention will be paid to the possibilities of using n-gram analysis and cluster analysis for taxonomic purposes. This line of study has been developed thanks to the cooperation with the University of Haifa in Israel, where the tradition of DNA Linguistics started decades ago and has been developing up to this day (BEREZOFSKY *et al.,* 2002; BOLSHOY, 2003). Methods from bioinformatics are used accompanied by tools from quantitative linguistics. The courses of DNA linguistics 1 and DNA linguistics 2 are taught by prof. Bolshoy from Haifa University. The Department of General Linguistics at Palacký University Olomouc also has its own tradition in analysis of genetic strings (FALTÝNEK *et al.,* 2019).

References

ALTMANN, G. (1978). Towards a theory of language. In: ALTMANN, G. (Ed.). *Glottometrika* 1. Bochum: Studienverlag Dr. N. Brockmeyer, 1978, p. 1-25.

AMARO, R. WordNet as a base lexicon model for the computation of verbal predicates. *Proceedings of GWC 2006*, Global WordNet Association Conference, 2006.

AMARO, R. Extracting semantic relations from Portuguese corpora using lexical-syntactic patterns. In: Conference LREC 2014 - 9th Language Resources and Evaluation, 2014, Reykjavik.

AMARO, R.; MENDES, S.; PALMIRA, M. Increasing Density through New Relations and PoS Encoding in WordNet.PT. *International Journal of Computational Linguistics and Applications*, v. 4, n. 1, p. 11-27, 2013.

BENEŠOVÁ, M.; FALTÝNEK, D.; ZÁMEČNÍK, L. Menzerath-Altmann Law in Differently Segmented Texts. In: BENEŠOVÁ, M.; MAČUTEK, J.; TUZZI, A. (Eds.) *Recent Contributions in Quantitative Linguistics*. Berlin: De Gruyter Mouton, 2015, p. 27–40.

BEREZOVSKY, I. N.; KIRZHNER, V. M.; KIRZHNER, A.; ROSENFELD, V. R.; TRIFONOV, E. N. Protein sequences yield a proteomic code. *Molecular Biology*, v. 36, n. 2, p. 239–243, 2002.

BOLSHOY, A. DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Applied bioinformatics*, v. 2. p. 103–112, 2003.

BUSA, R. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, v. 14, n. 2, p. 83-90, 1980.

CABRÉ, T. M. *Terminology: Theory, methods and applications*. Amsterdam: John Benjamins Publishing, 1999.

CALDARELLI, G. *Scale-Free Networks:* Complex Webs in Nature and Technology. Oxford University Press, 2007. Available in: https://EconPapers.repec.org/RePEc:oxp:obooks:9780199211517. Last accessed: 27 Apr. 2021.

CAMPBELL, R.; REIBLE, D.; TARABAN, R.; KIM, J. *More than a Dream:* The Developing Reflective Engineers through Artful Methods *(DREAM).* Project Paper presented at Gulf Southwest Section Conference, 2020. Avaible in: https://jee.org/36012. Last accessed: 27 Apr. 2021.

CHEN, R.; LIU, H. Thematic concentration as a discriminating feature of text types. *Journal of Quantitative Linguistics*, v. 25, n. 1, p. 53-76, 2018.

COLLIER, P. *The Future of Capitalism.* Facing the New Anxieties. London: Penguine Random House, 2018.

COSTA, R., SILVA. R. et al. *Glossário Colaborativo COVID-19.* Available in: https://www.lexonomy.eu/ec25mm79/ Last accessed: 27 Apr. 2021.

DERRIDA, J. *Of Grammatology.* London: Johns Hopkins University Press, 1976.

EUROPEAN COMMISSION. *Strategic Plan 2020-2024 – Education, Youth, Sport and Culture.* 2020. Available in: https://ec.europa.eu/info/publications/strategic-plans-2020-2024_en Last accessed: 27 Apr. 2021.

EUROPEAN COMMISSION. *Strategic Plan 2020-2024 – Research and Innovation*. 2020. Available in: https://ec.europa.eu/info/publications/strategic-plans-2020-2024_en Last accessed: 27 Apr. 2021.

EUROPEAN COMMISSION. *The Digital Europe Programme*. Available in: https://ec.europa.eu/digital-single-market/en/europe-investing-digital-digital-europe-programme Last accessed: 27 Apr. 2021.

FALTÝNEK, D.; MATLACH, V.; LACKOVÁ, Ľ. Bases are Not Letters: On the Analogy between the Genetic Code and Natural Language by Sequence Analysis. *Biosemiotics* 12, p. 289–304, 2019. Available in: https://doi.org/10.1007/s12304-019-09353-z Last accessed: 27 Apr. 2021.

FALTÝNEK, D., MATLACH, V. (2021). Forthcoming

FERRER-I-CANCHO, R., SOLÉ, R. V. The small world of human language. *Proceedings of the Royal Society B. London*, 268, p. 2261–2265, 2001.

FOUCAULT, M. *The Archaeology of Knowledge*. New York: Pantheon Books, 1972 [1969].

FOUCAULT, M. *The order of things:* An archaeology of the human sciences. New York: Vintage Books, 1994[1966].

GIERE, R. N. *Scientific Perspectivism*. Chicago: The University of Chicago Press, 2006.

GLOGAROVÁ, J. D., & KUBÁT, M. Srovnávací frekvenční analýza exilových projevů Klementa Gottwalda a Edvarda Beneše z let 1939-1945. *Slovo a Slovesnost*, v. 81, n. 1, p. 65-77, 2020.

GRZYBEK, P. Introductory Remarks: On the Science of Language in Light of the Language of Science. In: Grzybek, Peter (Ed.) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht: Springer, 2006, p. 1-14.

HERDAN, G. *The Advanced Theory of Language as Choice and Change*. Berlin: Springer-Verlag, 1966.

HŘEBÍČEK, L. Some aspects of Power Law. *Glottometrics*, v. 6, p. 1-8, 2003.

JANEČKOVÁ, B. A.; TICHÁ, A.; FIEDLER, J. *Třikrát o autorství*. Olomouc: Palacký University Press, 2021.

JENSEN, K. E. Linguistics in the digital humanities: (computational) corpus linguistics. MedieKultur: Journal of Media and Communication Research, v. 30, n. 57, p. 115-134, 2014. Available in: https://doi.org/10.7146/mediekultur.v30i57.15968 Last accessed: 27 Apr. 2021.

KELLERT, S. H. *Borrowed Knowledge: Chaos Theory and the Challenge of Learning across Disciplines*. Chicago: The University of Chicago Press, 2008.

KLEIN, N. *No Is Not Enough:* Resisting Trump's Shock Politics and Winning the World We Need. Toronto: Knopf Canada, 2017.

KÖHLER, R. *Quantitative Syntax Analysis*. Berlin: De Gruyter, 2012.

KÖHLER, R. *Synergetic Linguistics*. In: KÖHLER, R.; ALTMANN, G.; PIOTROWSKI, R. G. (Eds.) *Quantitative Linguistics:* An International Handbook. Berlin: Walter de Gruyter, 2005, p. 760–774.

KÖHLER, R. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer, 1986.

KOŘENSKÝ, J. Procesuální gramatika v kontextu současných tendencí lingvistického myšlení. *Slovo a Slovesnost*, v. 64, p. 1-7, 2003.

KUBÁT, M.; MATLACH, V.; ČECH, R. *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag, 2014.

LAUDAN, L. Dissecting the Holist Picture of Scientific Change. In: CURD, M., COVER, J. A. (Eds.) *Philosophy of Science:* The Central Issues. New York: W. W. Norton & Company, p. 139-169, 1998.

LEE, K.-F. *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston, Mass: Houghton Mifflin, 2018.

LESCH, K. *On the Linguistic Structure of Emotional Meaning in Czech*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 2015.

LIU, H.; LIANG, J. (Eds.). Motifs in Language and text. v. 71. Berlin: De Gruyter Mouton, 2017.

MEYER, P. Laws and Theories in Quantitative Linguistics. *Glottometrics*, v. 5, p. 62-80, 2002.

POPESCU, I. I.; MIANGAH, T. M.; GNATCHUK, H.; CECH, R.; BODOC, A; ALTMANN, G. On Rank-Frequency Distributions in Poetry. *Glottometrics*, v. 38, p. 30-54, 2017.

QUINE, W. V. O. *From a Logical Point of View*. New York: Harper, 1953.

ROSS, A. *The industries of the future*. New York: Simon & Schuster, 2016.

SALGADO, A.; COSTA, R. (2019). *LeXmart: a smart tool for lexicographers*. In: Electronic lexicography in the 21st century (eLex 2019): Smart lexicography, 2019, Sintra.

SALGADO, A.; COSTA, R. (2020). O projeto "Edição Digital dos Vocabulários da Academia das Ciências": o VOLP-1940. *Revista Da Associação Portuguesa De Linguística*, v. 7, p. 275-294, 2020. Available in: https://doi.org/10.26334/2183-9077/rapln7ano2020a17 Last accessed: 27 Apr. 2021.

SMITH, B. C. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge: The MIT Press, 2019.

VARMUŽOVÁ, B. Určování autorství Slezských písní. Olomouc, [cit. 2021-01-17]. Dostupné z: https://theses.cz/id/3p64qh

VEYNE, P. *Foucault*: His Thought, His Character. Cambridge: Polity Press, 2010.

ZÁVODNÍ, Š. Proměny verbálního projevu účastnic výchovně vzdělávacího procesu ve vztahu k biorytmům. Hradec Králové, 2020 [cit. 2021-01-17]. Dostupné z: https://theses.cz/id/divq2d/

Appendices

Appendix: Structure of the study plan of bachelor and doctoral programs in Linguistics and Digital Humanities (LDH) at Palacký University Olomouc

1 Bachelor program LDH: structure of the program

LDH can be studied either as a separate study program or in combination with another field of study.

In the case of an independent study program, there is a higher expectancy for the graduates to immediately transition into practice (with possible further study along the employment). In the case of combination with another study field, it is more likely that the student will follow a master's degree.

The independent study program LDH includes, in addition to the basic composition of the courses of the theoretical basis (TB) and profiling basis (PB), an elaboration of an individual project. If a student creates only one Independent Project, he/she must complete at least one Internship (10 credits). Individual projects are selected on the basis of consultation with the guarantor of the study cycle or its authorized representatives. A separate project ideally forms the basis for creation of a diploma thesis. The internship is arranged by the student with regard to their future employment.

In combination with another study field, LDH can be studied either as a *maior* or as a *minor*. For the *maior variant,* the student must complete the compulsory courses of the theoretical basis (TB, 54 credits) and profiling basis (PB, 18 credits). The student must also complete a diploma module (15 credits) and prepare a bachelor thesis in the field of LDH. The *minor* variant differs only in absence of the diploma module.

The three possible study programs can be schematized in terms of ECTS as follows:

Table 1: ECTS differences of bachelor LDH program in three possible varieties

Program	Independent	Maior 60%	Minor 40%
1st year	60	36	24
2nd year	60	36	24
3rd year	60	36	24
Total	180	108	72

The whole structure of the study program is schematized in table 2.

Table 2: Bachelor study program LDH

Course title	**ECTS**	**Year/ semester**	**profiling basis**
Theory of Humanities 1	5	1/WS	TB
Semiotics	6	1/WS	TB
Critical Discourse Analysis	5	1/WS	TB
Theory of Humanities 2	5	1/SS	TB
Forensic Linguistics	6	1/SS	TB
Introduction to General Linguistics	5	2/WS	TB
Data Processing in DH 1	5	2/WS	TB
Past of the Database Systems	5	2/SS	TB
Data Processing in DH 2	6	2/SS	TB
Present of the Database Systems	6	3/SS	TB
Alghoritminc Language Processing 1	4	1/SS	PB
Alghoritminc Language Processing 2	5	1/SS	PB
Alghoritminc Language Processing 3	4	2/WS	PB
Alghoritminc Language Processing 4	5	2/WS	PB
Diploma Thesis Topic	5	2/SS	
Diplomoma Seminar 1	5	3/WS	
Diploma Seminar 2	5	3/SS	
Individual Project 1	20	2/WS	PB
Individual Project 2	20	3/WS	PB
Internship 1	10	2/SS	PB
Internship 2	10	3/WS	PB
Aplied Semiotics	5	3/WS	PB
Text Atribution	5	2/WS	PB
Creative Visuality	5	1/SS	PB

Natural Language Processing	5	2/WS	PB
Mathematical Text Modelling 1	5	2/WS	PB
Mathematical Text Modelling 2	5	2/SS	PB
Mathematical Text Modelling 3	5	3/WS	PB
Science Methodlogy	4	1/WS	PB
Basis for Experimental Analysis of Language	6	1/SS	PB
Text Theory and Pragmatics	6	3/WS	PB
Introduction to Communication Theory	5	1/WS	PB
Psycholinguistics	4	–	PB
Biosemiotics	4	–	PB
Linguistic Applications	3	–	PB
Fiction and Reality Theory in Praxis	4	–	PB
Models of Linguistic Explanations	4	–	PB
Seminars of Invited Speakers	3	–	PB
Philosophy	3	1/WS	PB

2 Doctoral program LDH: structure of the program

The doctoral programs at Palacký University have a particularity of having ECTS system. Thus, every student has to acquire a particular number of credits in order to complete the study cycle. There are several modules (profiling, publications, teaching, foreign languages, etc.), each of the modules is characterized by a minimum of ECTS. The distribution of ECTS in every module is represented in Table 3.

Table 3: Doctoral study program LDH

Profiling Mandatory Courses	ECTS
Introduction to Digital Humanities 1 – Introduction to Quantitative Methods	10
Introduction to Digital Humanities 2 – Bases in DH: Text Processing and Multimedia	10
Philosophy of Science	5

Foreign Language	
To choose from the Faculty database of language courses	10

Profiling Optional Courses	
Linguistic Data Mining 1 – Data Analysis	10
Linguistic Data Mining 2 – Corpus Linguistics	10

Data Mining of Digitalised Text 1 – Introduction to Machine Learning 1	10
Data Mining of Digitalised Text 1 – Introduction to Machine Learning: NLP and Multimedia	10
DNA Linguistics 1	10
DNA Linguistics 2	10
Terminology and Organization of Knowledge	10
Presentation of Data and Access to Data	10
Python Programming	10
Biosemiotics	10
General Linguistics	10
Semiotic Approach to DH	10
Regulation of Cultural Industry and Digital European Market	10
Linguistic Applications	10
Linguistic Analysis of Historical Text: Application in History Studies and Germa Philology	10
Publication Activity	
Publication 1	10
Publication 2	20
Publication 3	30
Conference 1	5
Conference 2	10
Conference 3	20
Stay Abroad	
Stay Abroad over 30 days	20

Pedagogy Module	
Course teaching	5
Supervision of a bachelor thesis	5
Opponent to a bachelor or master thesis	3

Dissertation Module	
Quodlibet 1	5
Quodlibet 2	5
Dissertation Thesis Submission	60

According to the three profiling lines of the doctoral program, the profiling courses are:

In the specialization Linguistic data mining:
Forensic linguistics
Introduction to quantitative methods
Python programming

Data analysis
Data presentation and data access
Corpus linguistics
General Linguistics
Linguistic Applications

In the specialization of data mining of digitized text:
Linguistic analysis of historical texts – possibilities of use in German studies and History
Semiotic approach towards Digital Humanities
Introduction to machine learning
Introduction to machine learning: NLP and multimedia
Data analysis
Data presentation and data access
Python programming
Terminology and organization of knowledge
Regulation of cultural industries and the digital market in Europe

In the specialization of the DNA Linguistics:
DNA linguistics 1
DNA linguistics 2
Biosemiotics
Python programming

Submitted: 02/27/2021.
Accepted: 05/25/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 103-116, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p103-116

Artigo / Article

A deixis: uma proposta de anotação em XML no âmbito do texto

Deixis: A Proposal for XML Annotation within the Text

Miguel Magalhães*
migmaglit@fcsh.unl.pt
https://orcid.org/0000-0003-0055-8971

Matilde Gonçalves**
matilde.goncalves@fcsh.unl.pt
https://orcid.org/0000-0003-0039-4401

Resumo

Neste trabalho, propomos uma metodologia para a anotação dos deíticos e da deixis em XML. A utilização da linguagem XML para a anotação de *corpora* tem conhecido um desenvolvimento nos últimos anos, com a publicação de diversas metodologias ou refinamento de outras já existentes. Mas a anotação da deixis tem colocado problemas uma vez que esta opera a vários níveis. De facto, a análise da deixis depende de um conjunto de elementos linguísticos, mais ou menos expressos que não podem ser analisados individualmente, mas sim na relação que estabelecem entre eles e o contexto de produção e de circulação dos textos. Esta contingência conduz a problemas de sobreposição de níveis de análise. A metodologia, aqui apresentada, não só é sensível às relações com o contexto de produção e circulação dos textos, como também permite analisar essas mesmas relações sob diversas perspectivas.

Palavras-chave: Deixis; *Corpora*; Anotação; XML; Texto.

Abstract

In this work, we propose a methodology for annotating deictics and deixis in XML. The use of XML for corpora annotation has increased in recent years with the publication of several methodologies or the refinement of existing ones. However, noting deixis has posed problems since it operates on several levels. In fact, the analysis of deixis depends on a set of linguistic elements, more or less expressed, that cannot be analysed individually, but only within the relationship they establish between them and the context of production and

* Faculdade de Ciências Sociais Humanas da Universidade Nova de Lisboa; Centro de Linguística da Universidade Nova de Lisboa, Lisboa, Portugal. Bolsista da Fundação para a Ciência e Tecnologia (PD/BD/142789/2018).

** Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa; Investigadora do Centro de Linguística da Universidade Nova de Lisboa, Lisboa, Portugal.

circulation of texts. This contingency leads to problems with overlapping layers of analysis. This methodology is not only sensitive to these relationships with the context of production and circulation of texts, but also allows us to analyse these relationships from different perspectives.

Keywords: *Deixis; Corpora; Annotation; XML; Text.*

Introdução

O presente artigo[1] insere-se numa tese de doutoramento em curso, em linguística do texto e do discurso, e tem como objetivo apresentar uma proposta de anotação de *corpus* em XML que permita não só quantificar os elementos deíticos nos textos, mas também a visualização da relação que se estabelece entre estes elementos, nomeadamente temporais, espaciais e pessoais, na construção da deixis.

A anotação de *corpora* para o tratamento automático de línguas não é nova e existem algumas normas que têm sido publicadas e desenvolvidas (TEI, Eagles, a título de exemplo) mas estas normas, como aponta (HARDIE, 2014), são exaustivas e orientadas para projetos de anotação volumosos. Ou seja, não são necessariamente pertinentes para *corpus* de tamanho mais reduzido, nem podem ser executados por investigadores a um nível individual.

Atualmente, os trabalhos em linguística tendem a depender da análise de *corpus* construídos propositadamente, com anotações específicas e orientadas para o objeto de estudo em questão. Em consequência, assistimos, na última década, ao esforço de desenvolver metodologias e ferramentas que permitam analisar e observar de forma automática fenómenos semânticos em *corpora*.

No seguimento do trabalho que nos propomos realizar, no âmbito do doutoramento, a anotação dos deíticos é uma das ferramentas que nos permite perceber se existem padrões linguísticos associados à deixis e se esses padrões nos fornecem evidências para o estudo de géneros textuais. Deste modo, partindo das noções de parâmetros de género, mecanismos de realização textual e marcadores de género (MIRANDA, 2010), tal como foram definidos por (COUTINHO; MIRANDA, 2009), noções centrais na linguística do texto e do discurso, na linha do Interacionismo Sociodiscursivo (ISD) (BRONCKART, [1997] 1999), procuramos mostrar a relevância deste processo na aplicação prática nas áreas de ciência de dados. Concomitantemente, visamos perceber de que modo a ciência de dados pode funcionar como um instrumento auxiliar à análise textual de *corpora*.

Embora a deixis seja um dos aspetos a analisar, é também um dos mais complexos em termos técnicos pela própria natureza do objeto, como podemos observar:

[1] Trabalho financiado por fundos nacionais portugueses- Fundação para a Ciência e Tecnologia, como parte do projeto do Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

> [...] but now we are into deep theoretical water, because now the language of thought has indexicals, and in order to interpret THEM *we would need all the apparatus we employed to map contexts into propositions that we need in linguistics but now reproduced in the lingua mentalis, with a little homunculus doing all the metalinguistic work.* (LEVINSON, 2006, p. 100, sublinhado nosso).

Do que fica dito nesta citação de Levinson, a interpretação da deixis não está dependente apenas de uma descrição objetiva da linguagem mas do mapeamento dos contextos e de um conjunto de ferramentas que permitam descrever esse mapeamento. Assim, a análise da deixis depende de um conjunto de elementos linguísticos, mais ou menos expressos (como os tempos verbais, sujeitos expressos ou nulos, localizadores temporais e espaciais, entre outros) que não podem ser analisados individualmente, mas sim na relação que estabelecem entre eles e com o contexto de produção e de circulação dos textos.

Para além da introdução e da conclusão, este artigo estrutura-se em três partes. A primeira incide no tratamento automático de textos, a segunda sobre a deixis e a terceira sobre estrutura e anotação em XML.

1 *Corpora* e tratamento automático de textos

O uso de *corpus* para a análise linguística tem assistido, nos últimos anos, a um crescimento exponencial, ligado ao desenvolvimento de ferramentas e algoritmos, que permitem analisar, cada vez, maiores volumes de texto com ferramentas automáticas. A democratização do acesso a ferramentas informáticas tem permitido a constituição e uso de *corpus* anotados para trabalhos de investigação linguística mais específicos, permitindo construir *corpus ad-hoc*. A publicação de vários documentos normativos, como a *Text Encoding Iniciative* (TEI) ou *Expert Advisory Group on Language Engineering Standards* (Eagles) permitiu a estandarização da anotação de grandes *corpora*. No entanto, como referido em trabalhos anteriores (HARDIE, 2014), o formato TEI foi criado e desenvolvido num momento em que a criação de *corpora* era uma tarefa desenvolvida em projetos de investigação com equipas numerosas e, por isso, a anotação foi desenvolvida para ser profunda e para que tivesse o maior número de usos possíveis. Um dos exemplos desta prática é o British National *Corpus* (BNC) que foi desenvolvido sob a alçada de um consórcio composto por instituições públicas e privadas, e cujas equipas eram compostas por linguistas de várias áreas, bem como programadores informáticos. O BNC tornou-se, por isso, um *corpus* de referência para a investigação em linguística do inglês (britânico).

Atualmente, os *corpora* são construídos e anotados para projetos de investigação específicos, com equipas pequenas ou por investigadores individuais que não têm meios nem tempo para seguir as normas publicadas ou, porque não têm interesse de tornar público esse trabalho de anotação. Deste modo, têm surgido trabalhos que, não abandonando completamente os documentos normativos, propõem formas mais simplificadas de anotação (HARDIE, 2014). Por outro lado, o uso de *corpus* anotados que até recentemente se restringia à análise de

fenómenos lexicais, morfológicos ou sintáticos, tem-se alargado à análise semântica, como consequência do interesse no desenvolvimento de ferramentas para o processamento automático das línguas, em áreas como *data mining*, *text mining*, *marketing*, *machine learning*, entre outros.

A análise semântica dos textos, através de ferramentas automatizadas, coloca vários desafios, tanto da perspectiva linguística como da perspectiva técnica:

- A análise semântica tem uma componente subjetiva porque é uma avaliação que se faz sobre o enunciado e/ou o texto.
- Depende de fatores intra e extra linguísticos (contexto de comunicação, ancoragem espácio-temporal, operações de recuperação e ancoragem de informação) que ainda não são totalmente conhecidos e compreendidos;
- Não sendo possível estabelecer com clareza as operações de recuperação e ancoragem de informação, porque são de natureza abstrata, é difícil definir algoritmos que permitam analisar os textos de forma automática.

No entanto, algum trabalho tem sido desenvolvido nesta área e a análise semântica do texto só pode ser executada até um determinado ponto: extração de entidades e relações, desambiguação e análise de sentimentos, por exemplo. Mas o processamento das línguas naturais (NLP, sigla inglesa) a um nível mais profundo requer um conhecimento alargado do contexto de comunicação, e que é designado frequentemente por "senso comum". O senso comum refere-se ao contexto de produção e reconhecimento do uso do documento. É este contexto que permite a desambiguação do processo de comunicação:

> Reasoning about time is one of the most important aspects of commonsense reasoning. Linking a formal theory for time with an annotation scheme aimed at extracting rich temporal information from natural language text is significant for at least two reasons. It will allow us to use the multitude of temporal facts expressed in text as the ground propositions in a system for reasoning about temporal relations. It will also constitute a forcing function for developing the coverage of a temporal reasoning system, as we encounter phenomena not normally covered by such systems, such as complex descriptions of temporal aggregates. (HOBBS; PUSTEJOVSKY, 2003, p.74)

No que concerne aos estudos linguísticos relativos à análise do contexto, destacamos os trabalhos realizados no âmbito da Teoria Formal Enunciativa de A. Culioli (1999), sublinhando a compatibilidade epistemológica deste quadro com o do ISD. Sendo um modelo descritivo e explicativo sustentado em operações predicativas e enunciativas, possibilita dar conta dos diversos ajustamentos intersubjetivos e discursivo-textuais realizados pelos produtores textuais em função dos constrangimentos de origem contextual e sócio-cultural (CULIOLI 1981, p.53-54).

Esta teoria introduz um modelo descritivo e explicativo do modo como a significação é construída através e pela enunciação, evidenciando a importância do contexto na interpretação:

> [...] the context-dependence principle is considered as the decisive factor underlying any strict form of the principle of compositionality. According to this understanding, the study of natural languages is regarded as a study of interpretation. (VALENTIM, 2015, p.297)

De acordo com o exposto, fica claro que o utilizador não pode projetar exclusivamente as unidades linguísticas e as suas combinatórias no enunciado, tem, sim, de observar sistematicamente o contexto em que ocorrem (VALENTIM, 2015, p. 297). A vantagem desta abordagem é que (i) permite descrever a língua de uma forma mais complexa do que composicional e (ii) pressupõe que exista um conjunto de operações, comuns a um grupo, com as quais podemos reconstruir e interpretar os enunciados produzidos por outros. Qualquer tipo de anotação que pretenda descrever os aspetos semânticos de um texto tem, necessariamente, de relevar estes dois pontos: descrever o valor semântico dos elementos e tornar visível as operações de interpretação que lhe estão subjacentes, através da interrelação que se estabelece entre os elementos textuais.

Uma das marcas visíveis das operações cognitivas e linguísticas, elaboradas pelos enunciadores e coenunciadores e altamente dependentes do contexto, são os deíticos, que iremos abordar na seção seguinte.

2 Deixis

A relação entre discurso e contexto tem sido largamente debatida ao longo do tempo e não está no âmbito deste trabalho explanar as implicações teórico-filosóficas dessa relação. Vamos, por isso, partir da reflexão feita por Fernanda Fonseca e assumir que:

> A relação dinâmica que se estabelece entre o discurso e o contexto tem o seu fulcro nas operações de referenciação deítica, isto é, as operações que pressupõem a existência de um contexto referencial e viabilizam a sua representação conceptual sob a forma de um mundo. (FONSECA, 1992, p.138)

Deste modo, a construção dos referidos "mundos" está dependente do estabelecimento de coordenadas espácio-temporais e da rede de relações que estas estabelecem no ato enunciativo que as institui. Assim, os deíticos e localizadores surgem como a face visível desse mundo e a análise do seu funcionamento constitui uma via de acesso às operações enunciativas que permitem a construção do referente pela língua.

Destacamos igualmente o contributo de Lyons (1977) na definição da deixis:

> By deixis is meant the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatiotemporal context created and sustained by the act of utterance and the participation in it, typically, of a single speaker and at least one addressee. (LYONS, 1977, p.637)

Face ao exposto, assume-se que a deixis é o processo através do qual é feita a construção de valores linguísticos, que resultam na representação da referência, relativamente aos sujeitos

e ao espaço-tempo no enunciado e no texto. Esse processo é materializado linguisticamente através dos deíticos, enquanto "gestos verbais cuja função primária é estabelecer a ligação entre o explícito e o implícito na comunicação verbal". (FONSECA, 1992, p. 70)

Deste modo, a mostração ou ancoragem situacional que a deixis revela não pode ser observada e analisada apenas num campo físico, sensorialmente observável, mas num contexto que é partilhado: "a memória e os contextos de vária ordem a que podem implicitamente recorrer". (FONSECA, 1992, p.71)

Do que fica dito anteriormente, a deixis não se realiza apenas através das marcas que o locutor-enunciador lhe imprime, mas através de operações:

> Assim, as propriedades das diferentes formas e os valores das construções, tendo uma incidência na gramática de cada língua, constituem-se como marcadoras de operações abstratas, observáveis, nos textos através das diferentes "determinações" que desencadeiam (CORREIA; PEREIRA, 2015, p.51)

São estas marcas linguísticas com valor deítico que nos propomos a anotar e quantificar por forma a visualizar a relação que se estabelece entre estes elementos nos textos com o contexto.

2.1 Espaço, tempo e pessoa: a ancoragem deítica

Como referimos anteriormente, os deíticos permitem fazer a ancoragem do enunciado e interpretar quais são os sujeitos enunciadores e compreender as relações espácio-temporais que se estabelecem entre factos e referentes do ponto de vista dos interlocutores. Os deíticos são de três tipos: pessoais, temporais e espaciais.

Relativamente aos sujeitos da enunciação, as formas pessoais que adquirem valor deítico são as formas das primeiras e segundas pessoas, tanto no singular como no plural:

> First and second person are the only forms among all the pronominal forms mentioned above that have deictic values, since both function as subjective linguistic indices. In enunciative terms, the first person results from the process of identification that may occur between the enunciator subject/speaker and the enunciation subject (which coincides with the syntactic subject). In turn, the second person arises from the differentiation between the enunciator subject/speaker and the enunciation subject. However, in this case the enunciation subject (or syntactic subject) identifies itself with the co-enunciator subject/interlocutor. The third person does not assume any deictic value. (VALENTIM, 2015, p.300)

Deste modo, as marcas de pessoa, expressas pelos pronomes pessoais e marcas de flexão verbal (no caso do português europeu), pronomes/determinantes possessivos e pronomes de tratamento constituem-se como os elementos necessários ao nosso trabalho, enquanto marcas externas das operações cognitivas e linguísticas da referenciação.

Os deíticos espaciais são formas que indicam o espaço e que permitem a interpretação em relação ao espaço da enunciação. Assim, a deixis espacial inclui não só os advérbios e expressões adverbiais de lugar, mas também os pronomes e determinantes demonstrativos que indicam proximidade ou afastamento do locutor e/ou recetor, e também os verbos que demonstram movimento / localização de e para o espaço do emissor, como podemos observar no exemplo seguinte:

(1)

Chegámos à CES 2015, Consumer Electronics Show de Las Vegas, com alguma expectativa sobre os televisores quantum dot LED. O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado em alguns modelos da série Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório. (ID718)

No exemplo (1) podemos observar que as marcas morfológicas de 1ª pessoa do plural são um deítico pessoal e o verbo "chegar" com a preposição "à" indica um movimento a partir do espaço do interlocutor, funcionando, em interrelação com o evento (CES), como um localizador espacial. É nesta localização espacial que se irá desenvolver o enunciado. Este exemplo mostra também que a deixis não funciona exclusivamente com formas e construções fechadas e fixas, mas numa rede de interdependências, quer linguísticas, quer extralinguísticas, como a seguir evidenciamos.

No exemplo (1), podemos também observar a deixis temporal. O primeiro elemento a considerar é o tempo verbal pretérito perfeito simples (PPS) (*chegámos* e *confrontámos*) ao qual é atribuído um valor de anterioridade em relação ao momento da enunciação (T0). Concomitantemente, a presença de dois localizadores temporais autónomos (neste caso, a data *2015* e *2013*) irão determinar o ponto de partida das duas sequências temporais, a partir das quais se vai desdobrar a temporalidade semiotizada neste segmento textual: uma relativa ao eixo temporal de 2015 construída com o tempo verbal PPS e outra relativa ao eixo temporal de 2013 edificada com o mais-que-perfeito composto (*tinha sido* e *tínhamos testado*).

Este exemplo mostra como as formas linguísticas e os valores semânticos que estas assumem estão interdependentes e não são estáticas. O exemplo em análise mostra como a anotação morfosintática não é suficiente para dar conta das relações de interdependência que constroem as referências temporais e que são atualizadas através de operações de referenciação, a saber uma referência temporal relativa ao momento de enunciação (T0), outra relativa ao localizador 2015 e uma terceira a 2013, retomada anaforicamente pela expressão temporal "na altura". E que o valor dos verbos (valores de anterioridade, posterioridade e simultaneidade) é atualizado em função destes localizadores e da relação que se estabelece entre eles.

Qualquer anotação que queira dar conta destas operações tem de ter em conta três aspetos: i) a ancoragem com o momento de enunciação/produção do texto; ii) a coordenação entre os elementos de construção de referenciação temporal e iii) o valor que eles adquirem nessas operações.

Na seção seguinte, iremos mostrar que, através de uma linguagem de marcação flexível como o XML, é possível anotar e relevar estas operações.

3 Modelo(s) de anotação: abordagens e limites

Nos últimos anos têm-se desenvolvido diversos trabalhos de caráter experimental e exploratório para a anotação de elementos semânticos nos textos. A anotação da deixis ainda não foi objeto de trabalhos específicos de anotação. Seja pela sua natureza abstrata, seja pela necessidade de mapeamento das operações abstratas que lhe estão subjacentes, como já referimos, a anotação da deixis coloca dois problemas ao nível técnico: (i) a necessidade de ligar elementos com formas e funções distintas (a ancoragem temporal relativamente ao momento da enunciação, por exemplo) e (ii) o mapeamento dessas mesmas operações de um modo que possa ser observável. Neste sentido procurámos trabalhos que oferecessem propostas para a resolução destes desafios técnicos e, por isso, analisámos algumas propostas de anotação que, não incidindo especificamente sobre a deixis, oferecessem soluções aproximadas. De entre os vários trabalhos analisados (GOECKE, LIINGEN, METZING, STIIHRENHERG, 2010), encontrámos alguns estudos para a anotação de coreferências e relações anafóricas, como Recasens, Martí; Taulé (2007) e Recasens, Martí; Taulé (2007b), este último abordando o conceito de *bridging* e oferecendo uma proposta de anotação em vários níveis (*multilevel*). Embora sejam propostas orientadas para a resolução de anáforas, oferecem soluções flexíveis que podem ser adaptadas à deixis. De acordo com os objetivos do trabalho, os autores propõem uma metodologia de anotação que dê conta de dois tipos de ligações correferenciais: a deixis do discurso (*discourse deixis*) e as relações anafóricas. É importante esclarecer que a deixis do discurso é entendida e definida pelos autores como "as reference to a discourse segment, that is, to a non-nominal antecedente" (Recasens et al., 2007, p. 205), uma anáfora cuja referência é não-nominal, reservando as que têm como antecedente um NP ou VP para a anáfora propriamente dita:

> Our approach classifies bridging (or associative anaphors) those definite or demonstrative NPs that are interpreted on the grounds of a metonymic relationship with a previous NP or VP. (RECASENS et al, 2007, p. 205)

O uso do termo *deixis* afasta-se, portanto, da nossa definição. No entanto, este trabalho apresenta algumas vantagens, em relação a outras propostas, e que consideramos úteis e adaptáveis para o nosso objetivo. A primeira vantagem é que é orientado para a anotação da língua espanhola que, devido à proximidade com o português, lida com algumas idiossincrasias da língua (como as três formas dos pronomes demonstrativos) que não existem em inglês. A segunda vantagem, é o facto de os autores abordarem o texto enquanto uma unidade de sentido: "the text taken as a scene in the sense that it builds up both a textual and a contextual framework as the result of an interaction between the discourse and the global context". (RECASENS, et al, 2007, p. 206). Esta abordagem "global" ao texto aproxima-se da ideia de texto enquanto objeto comunicativo (BRONCKART, 1997) e objeto complexo (COUTINHO, 2003). No entanto, existe um outro ponto no qual o nosso trabalho diverge de Recasens et al. (2007) e que concerne à metodologia de análise. Se em Recasens et al (2007), a análise é metodologicamente ascendente, ou seja, parte das unidades linguísticas específicas para níveis mais amplos, a nossa metodologia análise sustenta-se numa abordagem descendente, na esteira dos trabalhos de

Bakhtine; Volochinov (1929/1977) e de Bronckart (1997/1999). Assim, a análise parte em primeiro das atividades de linguagem, passa pelos géneros de texto, enquanto formatos textuais formados pelas gerações anteriores, em seguida pelos textos enquanto materialização linguística dos géneros e objetos comunicativos complexos, para finalmente alcançar as unidades linguísticas que enformam os textos (BRONCKART, 1997, RASTIER, 2001, COUTINHO, 2003).

> Les textes sont des produits de l'activité humaine, et à ce titre (...) ils sont articulés aux besoins, aux intérêts et aux conditions de fonctionnement des formations sociales au sein desquelles ils sont produits. (BRONCKART, 1997, p.74).

Esta noção de texto implica que qualquer tipo de anotação, sobretudo da deixis, terá de ser sensível e, de certo modo, revelar esta ligação entre o contexto em que é produzido o texto e as unidades linguísticas presentes que edificam e revelam esta ligação.

O modelo de anotação que propomos, de acordo com a metodologia descendente, possibilita uma análise linguística que tem em conta a influência dos fatores físico-culturais que atuam na criação dos valores semânticos e discursivos adquiridos e estabilizados no texto.

Na seção seguinte, iremos fazer uma breve descrição do *corpus* utilizado neste trabalho e iremos mostrar alguns exemplos práticos.

3.2 O *corpus*

Como *corpus* de análise, foram utilizados textos selecionados, recolhidos no âmbito das atividades do grupo Gramática e Texto integradas no CoRus - Projeto Estratégico 2015-2020, desenvolvido no Centro de Linguística da Universidade Nova de Lisboa (CLUNL). Os textos dizem respeito a comentários e foram selecionados de acordo com a canonicidade, a representatividade e a atividade da linguagem em que se inseriram. Por canonicidade entendemos o estatuto de prestígio social, cultural, económico que uma determinada prática textual goza durante um certo período e que reflete uma legitimação consensual e normativa. A representatividade reflete-se nas várias formas de texto e não na representatividade estatística do mesmo. As atividades da linguagem em que se inserem os textos foram selecionadas pela representatividade que esta prática textual tem nestas áreas, nomeadamente, a jornalística, a académica e a jurídica. Os textos foram, depois, analisados pelo anotador categorial LX-Tagger e etiquetados em formato XML.

3.3 Estrutura e anotação

A anotação de um *corpus* permite tornar visível informação sobre o texto, seja informação de natureza implícita ou explícita (como autoria, data, fonte, entre outras) ou informação que está codificada na disposição física do texto (fronteiras de palavras, parágrafos ou frases, por exemplo), e para realizar estes processos são necessários tanto um processo conceptual como um processo técnico. De forma breve, consideramos o primeiro como o

modelo que parte de um conceito teórico (a sintaxe, a morfologia, a semântica, entre outras) e o segundo compreende a execução técnica da anotação, dependente do sistema de anotação escolhido para o processo. Assim, e para simplificar a terminologia usada neste trabalho, vamos designar o modelo conceptual de nível e a realização técnica de camada (GOECKE, LIINGEN, METZING; STIIHRENHERG, 2010).

A anotação foi organizada em várias camadas, cada uma englobando funções diferentes:

- Estrutural: engloba as anotações que definem a estrutura do documento (parágrafos, frases, corpo de texto, peritexto).
- *Inline annotation*: engloba anotações linguísticas que descrevem um elemento estrutural único (POS tag, lemmas, etc.)
- *Span annotation*: engloba anotações linguísticas que descrevem um ou mais elementos estruturais (entidades, expressões multipalavras, dependências, etc). Esta anotação está integrada num nível mais abrangente (como a frase ou parágrafo).

É nesta última camada que foram inseridas as marcações dos deíticos uma vez que a deixis é uma função da linguagem que se insere em diversas categorias gramaticais e opera ao nível (con)textual. Assim, a deixis foi incluída na *span annotation* <sa> como um *element* específico. Dentro deste *element* são incluídos os NP, VP ou SP, frase ou frases que tenham uma função deítica, como nos exemplos seguinte:

(1a)
No verão de 2015

(1b)
<sa function="deitic" type="time"> <w ort="EM"> <lex cat="PREP"></lex> No </w> <w ort="VERÃO"> <lex cat="CN"> </lex><msd gender="m" number="s"></msd> verão </w> <w ort="DE"> <lex cat="PREP"></lex> de </w> <w> <lex cat="DGT"> </lex> 2015 </w> </sa>

No exemplo (1b) podemos observar como através da função *element* podemos inserir os deíticos temporais da frase, neste caso introduzido pela preposição de tempo *Em*. A informação é de caráter descritivo com dois atributos: a função (*function*) e o tipo (*type*). Para a função definiram-se duas funções específicas, a deítica e a anafórica (embora não se enquadre neste trabalho, fica a referência para futuras anotações). Relativamente ao tipo de deítico, definiram-se três tipos de acordo com os deíticos possíveis: pessoais (*personal*), espaciais (*space*) e temporais (*time*). De notar que a proposta aqui apresentada é um exemplo específico que não contempla todos os elementos com função deítica, como os morfemas de tempo e de pessoa. Essa ancoragem só é possível se for feita dentro da descrição morfossintática do verbo. Nesse sentido, tornou-se necessário estabelecer uma ligação entre o elemento (morfológico ou outro) ao qual é atribuído o valor deítico e a anotação que descreve o valor deítico atribuído, e é nesta ligação que se estabelece a relação entre o nível e a camada. Definiu-se, por isso, dois atributos para o <sa>: o primeiro será *ref* e irá incluir um caminho *XPath* para o verbo em questão e o segundo *value* irá explicitar o valor que o deítico tem no texto em relação ao tempo verbal (posteridade, anterioridade, simultaneidade):

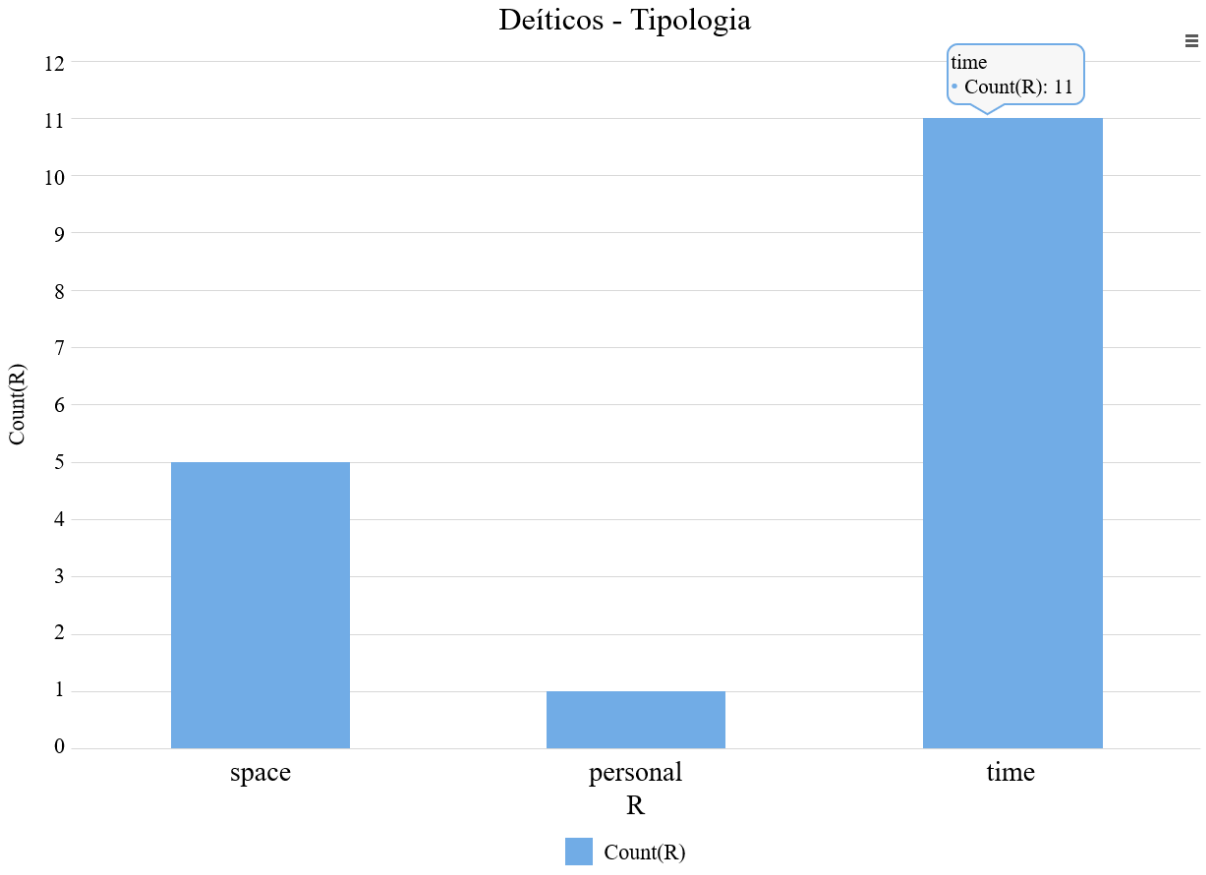
(2a)
Chegámos à CES 2015

(2b)
<sa function="deitic" type="space"> <w ort="CHEGAR"> <lex cat="V"></lex> <msd number="p" person="1" mood="ind" tense="pp"> </msd> Chegámos </w> <w ort="A"> <lex cat="PREP"></lex> à </w> <w ort="CES"> <lex cat="PNM"> </lex> CES </w> </sa>
<sa function="deitic" type="personal" ref="/Comentários[1]/comentário[1]/body[1]/p[1]/sent[1]/sa[1]/w[1]/msd[1]"></sa>
<sa function="deitic" type="time" ref="/Comentários[1]/comentário[1]/body[1]/p[1]/sent[1]/w[1]" value="ant"> <w> <lex cat="DGT"> </lex> 2015 </w></sa>

Do mesmo modo, os atributos *function* e *type* podem ser utilizados para os deíticos espaciais ("Chegámos à CES") e pessoais. Neste último, o objeto sobre o qual incidiu a <sa> foi a descrição morfosintática <msd> do verbo. Embora não seja visível, o XPath permite, neste exemplo, recuperar o atributo da <msd> que contém a informação sobre o tempo ou a pessoa.

Esta anotação que aqui se propõe permite recuperar os dados anotados, filtrá-los e analisá-los de várias formas:

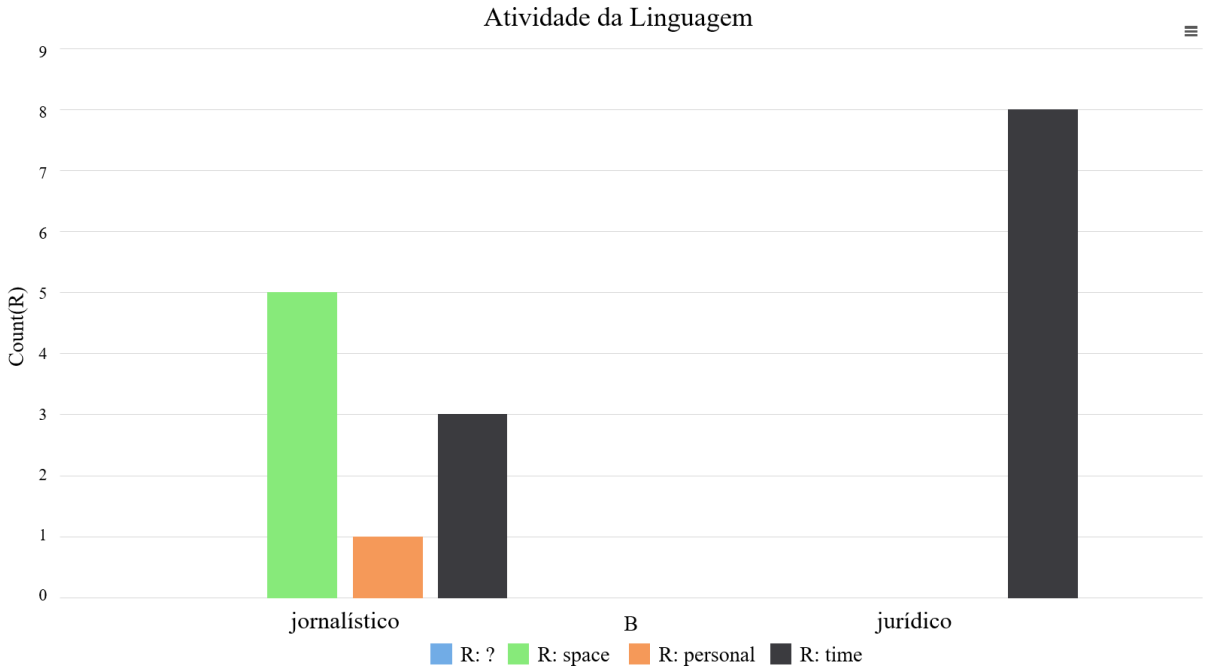
Figura1: Deíticos temporais



Fonte: elaborado pelos autores.

Na figura 1 podemos observar, por exemplo, que os deíticos temporais sobressaem nesta amostra. Mas, através de ferramentas de *text mining*, podemos também co-relacioná-los com a atividade da linguagem ou qualquer outra variável que estiver presente na anotação:

Figura 2: Relação entre deíticos e atividade de linguagem



Fonte: elaborado pelos autores.

Na figura 2, podemos observar a relação entre os diversos tipos de deíticos e a atividade da linguagem em que se inserem os textos (nesta amostra, a atividade jornalística e a atividade jurídica).

Conclusão

De caráter exploratório e experimental, o presente artigo visou apresentar uma proposta de anotação de *corpus* em XML possibilitando quantificar os deíticos (pessoais, temporais e espaciais) nos textos, bem como a evidenciação e a visualização das relações entre os diversos elementos da deixis e desses com as atividades de linguagem nas quais os textos circulam.

A deixis, sendo a face visível de um conjunto de operações abstratas, necessita de uma camada de anotação que opere em diferentes níveis de análise - morfológicos, sintáticos, semânticos e textuais. Esta característica da deixis evidencia como a categorização e etiquetagem de elementos semânticos e textuais coloca vários desafios técnicos que requerem um modelo flexível, capaz de atuar em diversos patamares (gramaticais e (con)textuais). Para tal, a anotação foi organizada em diversas camadas, com funções diferentes, agrupando nelas

diferentes funções e níveis de anotação, e que permitem recolher e visualizar informação sob vários ângulos e variáveis. No que toca à deixis, a marcação dos elementos deíticos realiza-se na última camada designada por *span annotation*. Esta camada engloba tipos de anotação (textual) que abrangem um ou mais elementos estruturais. A própria camada está embutida num nível estrutural de escopo mais amplo do que a palavra ou a frase, se necessário.

De destacar que a proposta de anotação, aqui apresentada, possibilita a visualização da sistematicidade dos valores que os deíticos adquirem no seio de um texto e de os analisar sob diversos ângulos. Estamos em crer que apresenta um grande potencial para colmatar a necessidade de criar ferramentas flexíveis atuando a níveis textuais meso e macro e em *corpus* menos extensos mas escaláveis. Sendo um trabalho exploratório, iremos continuar a desenvolver e a testar a metodologia, aumentando o número de textos e desenvolvendo novas soluções.

Referências

BAKHTINE, M., VOLOCHINOV, V. N. *Le marxisme et la philosophie du langage:* essai d'application de la méthode sociologique en linguistique. Préface de Roman Jakobson. Traduit du russe preésenté par Marina Yaguelo. Paris, Minuit, 1977.

BRONCKART, J.-P. Activité langagière, textes et discours. Pour un interactionisme socio-discursif. Paris: Delachaux et Niestlé, 1997.

CORREIA, C. N. ; PEREIRA, S. Formas e construções linguísticas no português europeu: Ferramentas referenciais e género textual. Cadernos de Linguagem e Sociedade, v. 16, n. 1, p. 48–60, 2015.

COUTINHO, M. A.; MIRANDA, F. To describe genres: Problems and strategies. In: BAZERMAN, C.; BONINI, A.; FIGUEIREDO, D. (Ed.). Genre in a Changing World. Fort Collins: The WAC Clearinghouse, 2009, p. 35-55. Available in: http://wac.colostate.edu/books/genre/ chapter3.pdf. Acesso em: 13 de nov. 2020.

COUTINHO, M. A. Texto(s) e competência textual. Lisboa: Fundação Calouste Gulbenkian-Fundação para a Ciência e a Tecnologia, 2003.

CULIOLI, A. Sur le concept de notion. Bulletin de Linguistique Appliqué e Générale, n. 8, p. 62-79, 1981.

CULIOLI, A. Pour une linguistique de l'énonciation. Formalisation et opérations de repérage (t2). Paris: Ophrys, 1999.

FONSECA, F. I. Deixis, tempo e narração. Porto: Fundação Eng. António de Almeida, 1992.

GOECKE, D.; LIINGEN, H.; METZING, D., STIIHRENHERG, M. Different Views on Mark up Distinguishing Levels and Layers. In: WITT, A; METZING, D. (Eds.), Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology. Dordrecht: Springer, 2010, p. 1-22.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. ICAME Journal, v. 38, n. 1, 2014, p. 73–103. Available in: https://doi.org/10.2478/icame-2014-0004

HOBBS, J. R.; PUSTEJOVSKY, J. Annotating and Reasoning about Time and Events. Working Papers of the 2003 {AAAI} Spring Symposium on Logical Formalization of Commonsense Reasoning, p. 74–82, 2003.

LEVINSON, S. C. The Handbook of pragmatics (L. R. Horn; G. Ward, Eds.). Choice Reviews Online, v. 41, Blackwell Publishing Ltd, 2006. Available in: https://doi.org/10.5860/choice.41-6349

LYONS, J. Deixis, Space and Time. In: STEINBERG, D. D.; JAKOBOVITS, D. D. (Eds.). Semantics. Cambridge: CUP, 1977, p. 636-724.

MIRANDA, F. Textos e géneros em diálogo: uma abordagem linguística da intertextualização. Lisboa: FCG/FCT, 2010.

RASTIER, F. Arts et Sciences du Texte. Paris: P.U.F, 2001.

RECASENS, M., MARTÍ, M. A.; TAULÉ, M. Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. International Conference Recent Advances in Natural Language Processing, RANLP, p. 504-509, 2007.

RECASENS, M., MARTÍ, M.; TAULÉ, M. Text as scene: discourse deixis and bridging relations. Procesamiento del Lenguaje Natural. Sociedad Española para el Procesamiento del Lenguaje Natural Jaén, n. 39, p. 205-212, 2007b.

VALENTIM, H. T. Deixis in European Portuguese: Representation and Reference Construction. In: JUNGBLUTH, K.; DA MILANO, F. (Ed.). Manual of Deixis in Romance Languages. Berlin/Bos: Mouton De Gruyter, 2015, p. 247-314.

Recebido: 28/02/2021.

Aprovado: 10/06/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 117-134, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p117-134

Artigo / Article

Interações digitais: conflito, argumentação e violência verbal nas redes sociais

Digital Interactions: Conflict, Argumentation, and Verbal Abuse on Social Media

Ana Lúcia Tinoco Cabral*
altinococabral@gmail.com
https://orcid.org/0000-0001-6417-2766

Manoel Francisco Guaranha**
manoel.guaranha@gmail.com
https://orcid.org/0000-0002-8676-601X

Resumo

Os comportamentos na Web constituem um vasto e frutífero campo de pesquisa em uma ampla gama de áreas científicas, especialmente a linguística. Interessamo-nos pelo comportamento linguístico dos usuários de redes sociais, principalmente do Facebook; investigamos a forma como as pessoas interagem, tomam partido em debates, argumentam, defendem pontos de vista, atuam em polêmicas. Partindo do questionamento se a grande visibilidade das interações nas redes sociais altera o estatuto da violência nas situações de controvérsia argumentativa, como é caso das polêmicas, investigamos como se dá o estatuto da violência nas redes sociais, especificamente, em interações marcadas pela dissensão, como é o caso das polêmicas. Nesse sentido, observamos como nessas interações prevalecem as paixões, o páthos, expresso especialmente pela violência verbal.

Palavras-chave: Discurso argumentativo; Polêmica; Discurso digital; (Im)polidez linguística; Discurso de emoções.

Abstract

Web performance is a vast and fruitful field for research in a wide range of scientific areas, especially linguistics. We are interested in the linguistic behaviour of social media users, particularly on Facebook; we investigate how people interact, take sides in debates, argue,

* Universidade de São Paulo - USP, São Paulo, Brasil; Faculdade De Filosofia, Letras e Ciências Humanas; Departamento de Letras Clássicas e Vernáculas; Mestrado Profissional / Pontifícia Universidade de São Paulo - PUCSP. pesquisadora colaboradora do IP-PUCSP, São Paulo, Brasil.

** Universidade Santo Amaro – UNISA, São Paulo, Brasil; Mestrado Interdisciplinar em Ciências Humanas/Faculdade de Tecnologia do Estado de São Paulo – FATEC, São Paulo, Brasil.

defend points of view, act in controversies. Starting from the question of whether the high visibility of interactions on social media changes the status of abuse in situations of argumentative controversy such as polemics, we investigate how abuse builds up on social media, particularly in interactions marked by dissent, namely controversies. In this sense, we observe how the passions, the pathos, especially expressed through verbal abuse, prevail in these interactions.

Keywords: *Argumentative Discourse; Polemics; Digital Discourse; Linguistic (Im)Politeness; Emotional Discourse.*

Considerações iniciais

Um dos fenômenos ligados às humanidades digitais de grande impacto nas relações de sociedade é o das redes sociais. Por meio delas, os indivíduos interagem com pessoas próximas e distantes, manifestam seus pontos de vista, resolvem problemas pessoais, atuam na sociedade. Os espaços de convivência migraram para os ambientes digitais, espaços que, conforme Amossy (2017), transformaram-se na praça pública do mundo contemporâneo, em que se discutem os problemas da sociedade. Atualmente, com as restrições de contato físico decorrentes do isolamento por conta da Pandemia de Covid-19, os ambientes digitais e as redes sociais passaram a ser os espaços primordiais de interação.

Somos seres sociais, e a sociedade é feita de seres variados, cujos valores e conhecimentos divergem. A sociedade inclui todos, com seus pontos de vista, sejam eles quais forem. Muitas manifestações nas redes sociais têm, no entanto, evidenciado comportamentos pautados na exclusão e na violência, que parecem se fazer mais visíveis no mundo contemporâneo, especialmente com as redes sociais. É fato que "a maioria dos atos de linguagem que são produzidos no cotidiano são potencialmente ameaçadores" (KERBRAT-ORECCHIONI, 2014, p. 49), a violência sempre esteve de alguma forma presente nas relações inter-humanas. O risco que constitui todo contato social e a "ideia de uma fragilidade intrínseca das interações" (KERBRAT-ORECCHIONI, 2017, p.18) conduziram à adoção de estratégias compensatórias, como os rituais de polidez. Não faltam, todavia, exemplos de situações nas quais a violência prevalece.

Os dispositivos digitais e as redes sociais possibilitaram maior visibilidade para as interações, para as pessoas, para suas manifestações, para as polêmicas e também para a violência. Essa maior visibilidade tem consequências sobre as quais este trabalho procura refletir.

Nesses ambientes digitais de redes sociais, os usuários criam perfis, ou seja, expõem sua identidade, que nem sempre é verdadeira. A identidade do usuário pode ser fictícia e, se não integralmente fictícia, é sempre possível construir um perfil idealizado, tanto para o bem como para o mal. O anonimato se torna, pois, possível por conta dos perfis fictícios. A respeito do anonimato, cumpre mencionar Cabral, Marquesi e Seara (2015), para quem o anonimato parece estimular o emprego de violência, pelo fato de o usuário, estando protegido pela falsa identidade

e pela máquina, não se sentir em risco. Na mesma direção de pensamento, Graham e Hardaker (2017) observam que esses dois fatores, o anonimato e o distanciamento físico garantido pelo computador, podem possibilitar que muitos usuários sejam mais sinceros e também agressivos. As observações desses estudiosos nos permitem inferir que a violência fica facilitada nas redes sociais.

Sendo as redes sociais espaço de discussões de temas de interesse da sociedade, muitas controvérsias se desenvolvem e inflamam as interações nesses ambientes, atraindo a atenção da pesquisa linguística interessada em compreender as interações, as argumentações e polêmicas que se desenvolvem nesses ambientes. Embora Amossy (2017), ao tratar da polêmica, afirme claramente que a violência não é um atributo obrigatório da polêmica, os processos de descrédito do outro ou de seus argumentos, próprios da polêmica, têm sido com frequência contemplados pela violência. Apenas para dimensionar a problemática, vale lembrar que, em 2020, houve, por parte de anunciantes poderosos, uma pressão sobre as plataformas de redes sociais para que houvesse controle da violência, e o *Facebook* anunciou medidas para contê-la. É sinal de que a questão da violência mais explícita nas redes sociais tem preocupado a sociedade.

Com base nessas breves considerações e assumindo que a violência continuará a existir, mas pensando no seu estatuto dentro das controvérsias argumentativas, em especial nas situações de polêmica nas redes sociais, uma pergunta orienta nossas análises e reflexões neste trabalho: a grande visibilidade das interações nas redes sociais altera o estatuto da violência nas situações de controvérsia argumentativa como é o caso das polêmicas?

O objetivo do trabalho é, pois, investigar como se dá o estatuto da violência nas redes sociais, especificamente em interações marcadas pela dissensão, como é o caso das polêmicas. Para tanto, o trabalho está organizado em quatro partes, além dessas considerações iniciais e das finais. Na primeira, abordamos as redes sociais como espaço de interação e construção de identidades; na segunda, tratamos teoricamente da polêmica como uma modalidade argumentativa (AMOSSY, 2017); na terceira, apresentamos algumas reflexões sobre a violência verbal e seu estatuto nas interações argumentativas; na quarta, analisamos a interação polêmica entre usuários do *Facebook,* observando a violência e o papel do *páthos*[1] nas interações escolhidas como *corpus*. As análises focalizam um exemplo singular de interação. São manifestações concretas que nos permitem apresentar algumas reflexões a respeito de interações verbais violentas nas redes sociais, encarando-as como eventos marcados pelo conflito e como fenômenos discursivos e argumentativos. Analisamos interações no perfil de uma revista cuja missão é propiciar aos usuários do *Facebook* uma reflexão de paz e harmonia. Contraditoriamente, comentários reativos a vários dos *posts* nesse perfil são marcados pela violência. Essa constatação conduziu-nos a questionamentos, a indagar o que leva os usuários à violência e a imaginar que, talvez, a violência seja um fenômeno necessário nas controvérsias.

[1] Seguimos a grafia preconizada por Houaiss e Villar (2001, p. 2149)

1 Redes sociais, interação e construção de identidades

Conforme expusemos anteriormente neste trabalho, as tecnologias digitais instituíram novas formas de interação e de atuação, e nesse contexto, as redes sociais ocupam um lugar de destaque, na medida em que propiciam aos seus usuários uma ampla convivência com pessoas do mundo todo. Criou-se, assim, por meio das tecnologias digitais, um novo modo de vida para a sociedade do século XXI. As formas de interagir propiciadas pelas redes sociais estabelecem igualmente novas formas de atuar no mundo. Conforme destacaram Cabral e Lima (2018, p. 40), a convivência entre os humanos, que antes se restringia "aos círculos sociais de vivência como trabalho, família, escola, esporte, igreja, entre outros, passaram também a comportar ambientes digitais, em plataformas de redes sociais, como o *Facebook*".

As redes sociais constituem um espaço no qual também se discutem os problemas de sociedade, sendo, conforme Develotte (2006), um ambiente de exposição discursiva, dado que os usuários interlocutores são produtores de discursos e estão constantemente expostos a discursos. As redes sociais se constroem de discursos em interação, constituindo um espaço no qual estão expostos sujeitos socialmente situados. Isso quer dizer que os usuários assumem uma posição nesses ambientes, situando-se em determinados espaços sociais, nos quais cada usuário tende a integrar determinado desempenho, ilustrando os valores que a sociedade reconhece oficialmente (GOFFMAN, [1974] 1981) ou pelo grupo com o qual se identifica. As interações nas redes são, pois, marcadas pelo fenômeno da idealização; trata-se de proporcionar aos interlocutores uma impressão idealizada, ou seja, uma identidade que o usuário considera ideal.

As plataformas de redes sociais utilizam termos ligados ao domínio da identidade, da construção do eu: perfil, amigos, centro de interesse, comunidade, o que, de acordo com Georges (2010), influencia para que os usuários percebam a comunicação como algo ligado ao *eu*. A esse respeito, vale lembrar Kaufmann (2004); esse autor observa que, nas redes, o usuário cria uma imagem que se instala e progressivamente constrói uma reputação, uma trajetória de vida. Essa instalação de imagem, reputação e identidade dá-se a partir das manifestações do usuário, ou seja, daquilo que ele diz em seus comentários ou dos *posts* que ele publica. Essa identidade está, muitas vezes, ligada a uma identidade de grupo.

Devemos considerar que, nas redes sociais, as pessoas selecionam seus grupos de interação com base em pontos de vista político-ideológicos ou por interesses comuns, o que acaba por criar comunidades de convivência por proximidade de interesses e pensamentos. Vale ressaltar que o algoritmo do *Facebook*, por exemplo, propicia esses encontros, por meio de sugestões de amizades e possibilidades de formação de grupos, propiciando a reunião dos usuários por suas preferências, o que cria uma identidade de grupo. Nas redes sociais, todos são *amigos,* e essas relações entre *amigos* fazem com que as hierarquias de poder sejam atenuadas, canceladas, pelo sentimento de pertencimento ao grupo de amigos, isto é, aqueles que pensam de forma semelhante. De acordo com Georges (2010), nos ambientes digitais, os usuários focam mais em seu desejo de expor uma identidade, o que os leva a se manifestar e a se tornarem mais pragmáticos; segundo o autor, esse fenômeno estimula e libera a imaginação.

Ainda relativamente à construção de identidades nas redes sociais, destacamos os postulados de Graham e Hardaker (2017), segundo os quais os usuários procuram ser sinceros, e a possibilidade de utilização de perfis fictícios facilita a sinceridade. Acreditamos que o emprego de violência pode estar ligado à sinceridade buscada, pois a linguagem violenta pode ser um meio de assumir um posicionamento frente a um conteúdo. Podemos afirmar que o usuário usa a violência para mostrar e argumentar em favor de uma identidade alinhada com determinado grupo. A violência também é facilitada pelos perfis fictícios, por detrás dos quais os usuários se escondem.

Observamos que, nas redes, criam-se comunidades de interação, nas quais todos são amigos. Amigos estão no mesmo nível hierárquico, o que permite maior descuido relativamente à linguagem empregada e também às regras de polidez, uma vez que, entre amigos, ficamos mais à vontade. Essa despreocupação pode conduzir à violência. Além disso, é importante considerar que os *posts* e as interações nas redes sociais se disseminam muito rapidamente, atingindo um vasto número de usuários, especialmente se pensarmos em perfis abertos e públicos. Essa ampliação constitui, de acordo com Paveau (2017), uma peculiaridade do discurso digital: há maior visibilidade para as interações, para os usuários e suas manifestações. Acreditamos que também as polêmicas que ocorrem nesses ambientes digitais atingem larga disseminação, e com elas, a violência nas discussões se espalha amplamente.

No Brasil, uma das redes de maior uso entre adultos é o *Facebook*, rede da qual é o terceiro país no mundo em número de usuários. As interações no *Facebook* ocorrem por meio de "*posts*, comentários aos posts e comentários reativos a comentários. Trata-se de uma tomada de posição com respeito a uma manifestação anteriormente enunciada por outro usuário que circula na rede" (CABRAL, 2019, p. 423). Esse processo põe em evidência a rede de interações em que se baseia o funcionamento da rede social *Facebook*, estabelecendo o diálogo entre usuários e posts e usuários entre si, propiciando a exposição de identidades em interação, o encontro de ideias e unindo pontos de vista, muitas vezes, radicalizados pelo estímulo de usuários de pensamento semelhante. A esse respeito, atribuindo um estatuto social discursivo ao comentário, considerando-o como um espaço social público, Seara e Cabral (2017, p. 314) observam que o comentário "permite a construção e a gestão da própria identidade".

2 Argumentação, Polêmica e redes sociais

Os estudos da argumentação costumam encarar o discurso argumentativo como aquele que busca o consenso em torno de uma questão. Assumimos, com Amossy (2017), que o consenso admite igualmente o dissenso, isto é, a diferença de formas de pensar e de julgar. O dissenso, segundo Amossy (2017), está na base da polêmica. É fato que os seres humanos não pensam todos da mesma forma, não julgam todos os fatos da mesma perspectiva nem com os mesmos valores, construindo, por consequência, avaliações diversas sobre um mesmo fato, julgamentos muitas vezes inconciliáveis. Vale ressaltar que, conforme observa Amossy (2017), o dissenso está na base da democracia, a qual admite que a sociedade ofereça respostas diversas para um mesmo problema.

Embora o acordo esteja no centro dos estudos da argumentação, a polêmica também constitui uma modalidade argumentativa, marcada pela oposição nos discursos; os debates e discussões marcados pelo desacordo são inclusive bastante frequentes na vida em sociedade. Amossy (2017) define a polêmica como "um debate em torno de uma questão de atualidade, de interesse público, que comporta os anseios da sociedade mais ou menos importantes numa dada cultura" (AMOSSY, 2017, p. 49). Em momento de pandemia, são inúmeros os debates em torno de temas variados associados à crise sanitária: é uma questão da atualidade, de interesse público e comporta anseios importantes de sociedade, e cada tema em discussão é mais ou menos importante dependendo da cultura.

Angenot (1982, p. 34) define que a "polêmica diz respeito a uma oposição de discursos no seio de um confronto verbal". A polêmica supõe, pois, um contra discurso antagônico, baseado na exposição de uma tese, sua demonstração e na refutação da tese contrária, muitas vezes desqualificando-a. Vale observar que Meyer (2008) encara a argumentação como um movimento voltado para uma solução de problemas, quando há respostas antagônicas para um mesmo problema, e cada uma das partes defende a sua proposta, apresentando seus fundamentos, sua argumentação. Para esse estudioso, a disputa argumentativa pode constituir a refutação de uma tese ou a produção de argumentos em favor de uma posição contrária, distinta. Esse postulado traz o contraditório como um princípio básico para a argumentação. Assim também ocorre com a polêmica.

É evidente que, para que haja polêmica, não basta declarar que é contra o argumento do outro; é preciso que a tomada de posição de um chame o outro a se manifestar para fazer valer suas razões refutando as do adversário. Além disso, cabe lembrar que a polêmica, segundo Amossy (2017), envolve um problema de sociedade, ou, como diz a autora, "que ela aborde um assunto de interesse público" (AMOSSY, 2017, p. 46) que suscite divergência de posicionamentos.

É por isso que a polêmica se situa no âmbito da argumentação. Não existe polêmica sem argumentação. Vale lembrar que Plantin (2016) define a argumentação como confronto dos pontos de vista antagônicos sobre uma mesma questão. Como lembra Amossy (2017), o próprio Plantin destaca que os traços definidores da polêmica e da argumentação parecem confundir-se. A polêmica é, pois, uma modalidade argumentativa marcada pelo choque das teses antagônicas (AMOSSY, 2017). Tomando a metáfora de Danblon (2005), quando ela aborda o panfleto, podemos afirmar que, assim como o panfleto, a polêmica provoca uma hipertrofia à crítica.

Nessa direção, Amossy (2017) ensina que a polêmica, verbalmente, caracteriza-se pela exacerbação das oposições, que conduz à dicotomização, uma vez que cada sujeito se mantém congelado em seu posicionamento. Esse comportamento divide os participantes da interação em grupos antagônicos. Essa situação de antagonismo hipertrofiado que leva à dicotomização conduz ao emprego de estratégias voltadas para o descrédito do opositor. Busca-se assim desqualificar o dizer do outro e a própria pessoa do adversário e levá-lo ao descrédito.

O descrédito do adversário tem por objetivo anular seus argumentos, afinal, se ele não é digno de crédito, seus argumentos também não o são.

É importante deixar claro que o caráter polarizado da polêmica, congelando os interlocutores em posicionamentos fixos, estabelece um funcionamento peculiar para a polêmica: o objetivo dos interlocutores não é persuadir o adversário a aderir às suas ideias, mas marcar um posicionamento relativamente aos outros. Assim, por exemplo, um usuário do *Facebook* que critica severamente aqueles que não tomam vacina contra Covid-19 mostra-se indignado com a negação à vacina para marcar que ele assume um posicionamento pró ciência contra aqueles que ignoram a ciência. Criam-se, desse modo, grupos de pessoas que partilham os mesmos pontos de vista, comunidades de interação.

Amossy (2017) observa que a opinião pública considera o *páthos* uma marca da própria da polêmica: "Essa opinião coloca a polêmica sob os auspícios da paixão em dois sentidos: como *páthos*, em seu sentido retórico, ou seja, como tentativa de suscitar afetos no auditório; e também como sentimento expresso com veemência por um locutor profundamente implicado na sua proposta" (AMOSSY, 2017, p. 137).

De acordo com os postulados de Amossy (2017), embora as emoções e a violência verbal estejam muitas vezes presentes na polêmica, elas não constituem atributos obrigatórios. Cabe observar, no entanto, que a violência verbal parece estar cada vez mais presente nas polêmicas, especialmente naquelas que se desenrolam nas redes sociais. Com respeito ao *páthos*, retomamos Amossy (2017, p. 138), "[a] questão da paixão tem importância na medida em que põe em causa a racionalidade da polêmica e sua capacidade de contribuir para a deliberação". Podemos dizer que é sob o signo do *páthos* que se desenrolam esses pseudodebates na Internet, pois "[s]e o raciocínio estiver animado pela paixão, é muito provável que esteja enviesado" (AMOSSY, p. 2017, p. 138).

3 Conflito de opiniões e violência verbal

Os seres humanos, com o intuito de se protegerem dos riscos que representam os encontros sociais, procuram formas de preservar a harmonia empregando estratégias de polidez (KERBRAT-ORECCHIONI, 2005). As redes sociais, no entanto, têm sido "palco para discussões violentas, de fortes controvérsias nas quais imperam discursos agressivos, cujos movimentos são marcados pela desqualificação do outro" (CABRAL, 2020, p.53).

O emprego de violência verbal, de acordo com autores como Culpeper (2008, 2011) e Bousfield (2008), é intencional, isto é, quem usa uma expressão violenta o faz com a intenção de agredir, mas o objetivo pode ir além da agressão simplesmente. Pode, por exemplo, servir para marcar um posicionamento hirto diante de uma questão em discussão ou para mostrar pertencimento a um grupo, uma comunidade caracterizada pela violência pela firmeza de seus pontos de vista. Como se o autor da violência raciocinasse no sentido de que, se todos no grupo são violentos, ele também o será para alinhar-se ao grupo. Sendo assim, nas redes sociais, a

violência funciona tanto como estratégia argumentativa quanto como elemento de construção de identidades. Desse ponto de vista, podemos inferir que comportamentos verbais violentos em interações nas redes sociais são passíveis de estimular a expressão de violência mais intensificada.

As expressões violentas desqualificam o interlocutor e o levam ao descrédito; lembramos que o descrédito do adversário constitui uma das peculiaridades da polêmica conforme Amossy (2017). Houaiss e Villar (2001, p 1629) expõem que o insulto expressa a "aversão ou menosprezo pelos valores, pela capacidade, inteligência ou direito dos demais". Essa definição do dicionário nos permite avaliar a extensão do dano causado pela violência, pois ela leva o adversário ao descrédito e seus argumentos perdem valor. Cabral e Albert (2017, p. 278), inclusive, observam que "o insulto não se limita a um ato verbal que agride" o outro; para essas autoras, a desqualificação abrange também o domínio social e prejudica a imagem do insultado. Ocorre que, na polêmica, o objetivo não é convencer o adversário, mas marcar a diferença, por isso é que os interlocutores se congelam em seus posicionamentos (AMOSSY, 2017); para essa autora, a polêmica, efetivamente, põe dois polos em interação, confrontando-se.

Nesse contexto de confronto, a violência parece ser prevista, afinal, como afirma Kerbrat-Orecchini (2014, p. 47), a "polidez nunca possui um lugar nas guerras, onde se trata, antes de tudo, de atacar o adversário para vencê-la, e assim também acontece nas guerras metafóricas que são os debates". A polêmica constitui um debate público sobre uma questão de sociedade, trata-se de uma guerra de pontos de vista antagônicos, no âmbito da qual, muitas vezes, vigora a violência, que, conforme Cabral e Lima (2017), precisa estar linguisticamente marcada. As autoras citam como exemplos de marcas de violência os qualificadores de caráter pejorativo.

Refletindo sobre o estatuto da violência nas polêmicas que ocorrem nas redes sociais, somos levados aos postulados de Mills (2017). Tratando do fenômeno da (im)polidez de um ponto de vista sociocultural, essa estudiosa destaca que os interlocutores seguem normas que eles assumem como adequadas a contextos sociais específicos. De fato, Culpeper (2011) ensina que a avaliação de um comportamento como impolido tem a ver com uma avaliação de que determinado comportamento é conflitante em relação ao esperado. Consideramos importante refletir sobre a relação entre polidez, violência e contexto, associando também às comunidades de redes sociais. Locher (2012) também propõe observar as questões de (im)polidez de um ponto de vista sociocultural, focando na relação entre os participantes da interação e nos efeitos que uma linguagem violenta, por exemplo, pode ter sobre os demais interlocutores. Sua preocupação é compreender o fenômeno em contextos situados. De fato, há contextos em que a violência é até esperada, e seus efeitos são outros; nas batalhas MC, como mostraram Oliveira e Cabral (2020), a violência é não apenas prevista, mas ela é também a regra, inclusive para ganhar a batalha. Neste artigo, análises procuram observar como acontecem as manifestações de violência verbal em um caso concreto de interação polêmica na rede social *Facebook*, visando verificar o estatuto da violência nesse contexto.

4 Polêmica, Argumentação e Violência na rede social *Facebook*

Nossas análises focalizam um *post* em um perfil de uma revista no *Facebook* e alguns comentários que compõem uma interação conflituosa a respeito do *post.* Antes de passar às análises propriamente ditas, cumpre esclarecer decisões que envolvem a coleta do *corpus*: Graham e Hardaker (2017) esclarecem, relativamente a pesquisas envolvendo a comunicação digital, que as pessoas, quando postam em fóruns *online* públicos, não podem ter expectativa de que seu comportamento não será examinado. |Esses estudiosos dos discursos digitais observam que os participantes de interações *online* em ambientes públicos, ao se manifestarem em ambientes aos quais outros terão acesso, aceitam implicitamente todas as consequências desse ato. Entendemos com Graham e Hardaker (2017) que o discurso público deve estar ao alcance do estudo científico. Apagamos todas as identificações dos usuários e utilizamos a proposta de Maíz-Arévalo (2019), de tomar precauções no sentido de preservar a identidade dos usuários. Dessa forma, mesmo sendo o *Facebook* um ambiente público, e a *Revista Pazes,* uma página aberta, pública, transcrevemos as mensagens sem identificação dos usuários. Utilizando procedimento semelhante ao adotado por Maíz-Arévalo (2019), escolhemos U1, U2, U3, para usuário 1, usuário 2, usuário 3 sucessivamente, por ordem de comentário, e acrescentamos a "U" a desinência de masculino "o" ou feminino "a", conforme o nome original dos usuários produtores dos comentários. Tendo esclarecido as questões de ordem ética e exposto os procedimentos de marcação dos comentários, passamos à exposição e análise do *corpus*.

O *corpus* analisado foi coletado em uma página institucional no *Facebook* da *Revista Pazes*[2]. O perfil dessa revista tinha, em 8 de março de 2021, data da coleta dos dados, 1.585.144 seguidores. Conforme declarado em seu perfil, a *Pazes* define-se como um site de entretenimento que tem como missão *"*levar, dia e noite, uma reflexão de paz, empatia, serenidade e equilíbrio a cada um dos nossos leitores" e assume que "a paz é sempre plural".

Podemos inferir que, pelo seu nome e sua missão explicitada na página inicial, a revista chama para a interação usuários interessados em paz, serenidade, equilíbrio e empatia. O quadro institucional da revista estabelece, pois, um quadro enunciativo de harmonia, a qual também deve prevalecer na linguagem empregada pelos usuários que seguem essa revista e participam das interações que nela ocorrem. Chamam a atenção, contudo, as manifestações violentas, que contrariam a missão da revista, provocando questionamentos relativamente à pluralidade da paz que a revista preconiza.

No dia 26 de fevereiro, a *Revista Pazes* publicou em seu perfil um *post* contendo a notícia do falecimento de uma enfermeira. O *post* ficou disponível até o dia 10 de março e suscitou muitas reações, considerando o número de comentários: 474, ressaltando que estão

[2] Disponível em: https://www.revistapazes.com. Acesso em: 08 mar. 2021

computados apenas os comentários considerados relevantes pelo sistema da plataforma do *Facebook*. O *post* apresenta o seguinte texto: "sim, meus amigos... Duvidar da ciência pode acarretar perda da própria vida! Nossos sentimentos à família". Em seguida a esse texto, o *post* apresenta a foto de uma jovem sorridente com a seguinte legenda: "enfermeira se recusa a tomar CoronaVac e falece de reinfecção da covid-19".

A notícia traz um dado de constatação expresso por dois fatos, entre os quais se estabeleceu um vínculo causal por meio do conector "e":

> Fato 1 - Enfermeira se recusa a tomar CoronaVac
> e
> Fato 2 – falece por reinfecção da covid-19

Essa constatação permite ao enunciador que assume a voz da revista emitir uma avaliação com valor de alerta: "duvidar da ciência pode acarretar a perda da própria vida", o que equivale a dizer aos usuários: "não duvide da ciência, você pode morrer". Verdade à parte, a mensagem tem um tom ameaçador, reforçado pela saudação "Sim, meus amigos". É a voz do saber que alerta.

Mesmo com um tom de conselho e ameaça, não se pode dizer que a mensagem seja, em si, agressiva ou violenta. Mas também não se pode dizer que a revista transmite harmonia nesse caso. A notícia poderia vir acompanhada de um alerta sobre a importância da vacinação, mas dito de forma que, de fato, transmitisse harmonia e estimulasse a vacinação. Violento é o fato narrado, a morte da enfermeira. A aditiva que conecta os fatos 1 e 2 sugere, para além da soma dos eventos, uma relação direta de causa e consequência entre eles. Ainda que suavizado pela mensagem cortês de "Nossos sentimentos à família", o caso da enfermeira serve como um exemplo e como uma provocação do enunciador àqueles que, eventualmente, tenham alguma dúvida sobre se devem tomar ou não a vacina. O que seria uma possibilidade, uma hipótese, a de poder perder a vida em caso de duvidar da ciência, concretizou-se para a personagem exposta no *post*. Desse modo, a revista toma partido na polêmica que se instaurou quanto à eficácia das vacinas contra o Coronavírus e, ainda que a posição seja legítima, pois a vacinação tem se mostrado eficaz no controle da Pandemia, o uso do exemplo extremo da morte da jovem, com certeza, apela para o *páthos*.

A morte de uma jovem enfermeira, especialmente de Covid-19, deveria mobilizar sentimentos de solidariedade. A morte em si causa um sentimento de dor e de solidariedade. Quando é uma pessoa jovem, que ainda teria muito a viver, esse sentimento normalmente é mais forte ainda. Assim seria esperado que as pessoas manifestassem empatia pela jovem e dor pelo desaparecimento dela. Essa empatia, contudo, desaparece face à provocação ativada pelo comentário da revista, e o objeto da inicial da postagem em si fica fora de questão em favor da polêmica entre os favoráveis e os desfavoráveis à vacina.

Nossas análises partem de um comentário inicial que desencadeou uma discussão. Esse conjunto de comentários nos permitem observar como se desenrola uma polêmica e como ela

se dissolve na paixão, no *páthos*. Os limites de um artigo acadêmico não nos permitem explorar todos os comentários que o *post* suscitou, mas é possível avaliar uma discussão. Iniciamos pelo comentário que desencadeou a discussão que tomamos para análise.

> Uo1 – Só pra lembrar quem tá morrendo de peninha dessa vítima, ela contraiu a doença e provavelmente passou para outras pessoas por pura imperícia, pois trata-se de alguém com conhecimento técnico que escolheu negar o que aprendeu na sua formação, é ridículo.

O comentário de Uo1 faz referência ao sentimento de solidariedade expresso em comentários anteriores e permite inferir que houve manifestações de pesar pela morte da jovem: "quem está morrendo de peninha dessa vítima". O emprego do pronome "quem" constitui uma construção de caráter exclusivo, isto é, exclui o locutor do fato enunciado e implica que Uo1 não está com pena da vítima. Há uma argumentação para justificar essa não inclusão, com dados de probabilidade vinculados ao saber científico relativo às possibilidades de transmissão do vírus. Sendo ela uma enfermeira, é de se esperar que detivesse tais conhecimentos: "Ela tem conhecimento técnico e escolheu negar". Uma adjetivação de valor negativo fecha a argumentação; trata-se de uma remissão encapsuladora que define todo o acontecimento como "ridículo", adjetivo extensivo aqui à própria vítima. Há várias marcas de violência verbal neste comentário, a começar pela expressão "morrendo de peninha", que desqualifica os sentimentos de quem tenha sido solidário à enfermeira "imperita", a qual teria escolhido o negacionismo e que, além de tudo, contaminou outras pessoas. Na visão de Uo1, a morte representou, portanto, merecido castigo para a enfermeira, uma espécie de pena capital. Essa provocação seguida àquela que identificamos na postagem da revista irá gerar outros comentários os quais irão desencadear a polêmica que se ampliará para além do drama individual da enfermeira e de sua família, para um problema da sociedade: tomar ou não a vacina, polêmica essa que está contaminada pelo embate político entre duas frentes ideológicas que têm sido marca da sociedade brasileira nos últimos anos. Essa questão opõe dois grupos em torno da polêmica: negacionista/ pró-vacina; eficácia de outros remédios que não a vacina contra a Covid-19/não eficácia desses remédios. O primeiro grupo tem se apoiado nos discursos do Presidente da República e de sua ala ideológica, enquanto o segundo tem se apoiado tanto em especialistas na área da saúde quanto utilizado o discurso desses especialistas para desqualificar o Presidente e seus seguidores. Essa discussão sanitária-político-ideológica tem, por assim dizer, contaminado a polêmica em torno dos temas relativos à Pandemia, especialmente nas redes sociais, e nosso *corpus* é uma mostra disso.

Apresentamos a seguir um excerto de interação em torno dessa polêmica. Embora não tenhamos todas as mensagens, porque algumas são filtradas pela plataforma, a sequência interativa permite-nos observar como se desenvolve a polêmica; com base nessa observação, procuramos verificar o estatuto da violência verbal na interação.

> Ua2 - Isso é prova de que a ignorância mata muito mais do que o vírus.
> Uo3 – Ua2 maior perigo é ficar como vc! Fuja louco!
> Uo4 - sério? Então pra que tomar a vacina se ela não funciona?? Ou você toma vacina contra rubeula (*sic*) e pega a doença?? Vacina contra pólio, contra paralisia, sarampo entre todas que funcionam vc já viu alguém pergar (*sic*) depois de tomar?
> Uo5 - Uo4 Não tome a vacina, cara. Mas não torra a paciência com a tua asneira.
> Uo4 – Uo5 asno é seu pai, estou só passando dados parem de pensar como asnos.
> Uo3 - Bando de hipócritas! Primeiro porque a vaChina seria preventiva e só funciona depois da segunda dose (ainda assim é no mínimo duvidosa), então ela jamais seria salva pela vaChina, o que está salvando vidas são os medicamentos que é a Hidroxicloroquina, Azitromicina e Zinco, ou Ivermectina e Nitazoxanida.. e outros que seguem quando o quadro vai para o estágio 3 que é mais grave.
> Ua6 – Uo3 e por que os medicamentos não salvou ela?????
> Uo3 – Ua6 pergunta ao médico o que ele usou?
> Uo7 – Ua6 porque esses medicamentos não são para covid.
> Uo7 – Uo3 ivermectina é para tratar os vermes do seu cérebro, né bolsominion?
> Uo3 – Uo7 pode ser tbm, quem sabe! Já quem tem cérebro de jumento aí mesmo que não adianta! Hehehe
> Uo8 – Uo3, que interessante! Ainda tem gente que acredita em hidorxicloroquina e ivermectina.
> Uo3 – Uo8, informação faz bem e não faz passar vergonha! Leia e ouça notícias de verdade e menos globo, isso vai lhe ajudar!
> Ua9 – Eitaaaa!!! A jumentada do lularápio encantador de burros, tão que tão!!!

O comentário de Ua2 associa não vacinar à ignorância e estabelece a recusa de tomar vacina como sendo causa de morte mais forte do que o vírus. O falecimento da jovem, segundo Ua2, é prova de seu ponto de vista. Instaura-se assim um ponto de vista pró-vacina, ao qual alguns usuários se oporão, enquanto outros o apoiam. A resposta de Uo3 indica que esse usuário discorda de Ua2, discordância exposta de forma violenta, qualificando o interlocutor de "louco". Trata-se de um adjetivo axiológico de valor negativo, que desqualifica a interlocutora, sugerindo que ela não merece crédito, uma vez que ela teria alterações patológicas nas faculdades mentais.

Vale lembrar que a busca da desqualificação do interlocutor no embate discursivo, em ambos os lados, compõe-se de palavras do campo semântico da irracionalidade: "ignorância", "loucura", "asneira", "vermes no cérebro", "cérebro de jumento", "lularápio encantador de burros" (esse último evidencia de forma mais clara a politização do tema a que nos referimos). É bastante significativo ambos os lados reivindicarem o território da razão atribuindo ao outro o da irracionalidade, ao tratar de um tema que, do ponto de vista científico, está bastante pacificado: a Covid-19 é uma doença infectocontagiosa cuja prevenção é feita pela vacinação prática médica que tem o fim de provocar a formação de anticorpos contra os agentes infectantes. No caso do *post*, a vacina citada é a Coronavac, que tem autorização temporária para uso emergencial pela ANVISA, Agência Nacional de Vigilância Sanitária, desde janeiro de 2021, além de ser usada em outros países. Em fevereiro de 2021, mês das postagens, cinco países, além do Brasil, já haviam aprovado o uso emergencial do medicamento: Indonésia, Turquia, Chile, Colômbia, Uruguai e Laos. Além disso, à época das postagens, a vacina havia conseguido o registro definitivo na China (CNN BRASIL, 2021). Nenhum dos interlocutores,

contudo, serve-se de argumentos técnicos com fontes. A violência assume o estatuto de uma espécie de bate-boca virtual em que pessoas adultas, a julgar pelas fotos das postagens, parecem estar em uma arena em que o que falta a todos é, justamente, aquilo que acusam que falta ao outro: racionalidade, que é a base da argumentação. Trata-se, aqui, da prevalência clara do *páthos* sobre a razão.

Nesse sentido, o embate discursivo amplia-se como uma conversa de pessoas com ouvidos moucos, a ponto de dois indivíduos com o mesmo ponto de vista ofenderem-se reciprocamente, como é o caso de Uo4 e Uo5. Em primeiro lugar, Uo4 questiona, de forma um tanto ambígua, aqueles que são contra a vacina, com uma lista de exemplos de outras vacinas que as pessoas tomam, estabelecendo uma analogia: se esta vacina não funciona, aquelas também não devem funcionar; e por que as pessoas tomam? O apoio na analogia serviria de argumento em favor de tomar a vacina. Como os dados da argumentação vêm em forma de pergunta, por meio de uma analogia e não por afirmação, não fica claro exatamente qual é a resposta que ele espera a essa pergunta. A ambiguidade é reforçada pela ironia na pergunta inicial "Sério?", sugerindo que o comentário de Uo3 não deve ser levado a sério, e, portanto, também não seu ponto de vista expresso no comentário.

Como resposta, Uo5 expressa uma crítica àqueles que são contra a vacina na forma de dois atos de ordem: 1. Não tome vacina, cara. 2. Mas não torra a paciência com a tua asneira. O vocativo "cara" e a atribuição de "asneira" à opinião de Uo4 equivale, na construção, a uma ordem para o outro calar a boca, o que encaminha a discussão, mais uma vez, para a violência, ao empregar linguagem agressiva, uma vez que o termo "asneira" encapsula todo o dizer de Uo4. O substantivo asneira remete a "asno", que está na origem do nome e designa um "dito impróprio, impensado". Como a palavra "asno" em sentido figurado designa um ser desprovido de inteligência, pode-se inferir que o Uo5 chama Uo4 de "burro". Aqui configura-se aquele diálogo entre surdos que destacamos, pois vale observar que o usuário Uo5 provavelmente não compreendeu bem o que o interlocutor defendia, já que ambos estão a favor do mesmo argumento: tomar a vacina. A reação de Uo5 é, contudo, de quem compreendeu o oposto.

Como reação, Uo4 parte para a agressão, com violência verbal explícita: "Asno é seu pai, estou só passando dados parem de pensar como asnos". A palavra "asno" é empregada duas vezes nesse curto comentário. A violência é empregada para fortalecer o posicionamento a favor da vacina, mas cabe observar que, em sua defesa, Uo4 afirma fazer justamente o que não fez: passar dados, que seriam resultados de investigação, de pesquisa ou conhecimento prévio para se resolver um problema. O sujeito, desse modo, reivindica credibilidade para suas impressões, apoiado nos supostos dados que afirma repassar, sem, contudo, dizer quais são, de que natureza são ou de que fonte os colheu. A pálida referência aos dados assume aquele caráter tão comum nessas discussões que é a de apropriar-se de argumentos pseudocientíficos ou supostamente científicos. Em alguns casos, nessas discussões, aparecem vagas referências a estudos científicos, pareceres de especialistas, entre outras construções que, curiosamente, servem-se da ciência para desconstruir o próprio conhecimento científico. Ainda que Uo4 esteja

certo em seu posicionamento a favor da vacina, caso seja esse mesmo, dada a ambiguidade detectada em seu comentário, tanto ele quanto os outros que se posicionam da mesma forma acabam prestando pouco serviço à causa defendida, movidos que se encontram pela paixão. Na sequência da interação, surgem manifestações cujo posicionamento é contrário à vacina e em defesa de medicamentos como cloroquina, entre outros, tão enviesados pela paixão quanto os anteriores.

O usuário Uo3 dá continuidade à discussão, iniciando seu comentário com uma violência explícita, na forma de um xingamento: "bando de hipócritas". Bando remete a bandidagem, grupo de malfeitores. Somente essa designação já traz um valor negativo; esse grupo de malfeitores é, em adição, caracterizado pela hipocrisia, ou seja, pela dissimulação, pela falsidade, pela mentira. A argumentação para justificar a desqualificação vem da crença de que a vacina não funciona contra a doença. Uo3 argumenta que o que funciona são medicamentos que ele lista, cinco, além de outros, conforme Uo3. O comentário de Uo3 contrário à vacina gera uma discussão que o coloca em questionamento e refutação da afirmação em defesa dos medicamentos, e Uo7 apresenta um comentário, dirigido a Uo3, marcado explicitamente de violência: "Ivermectina é para os vermes do seu cérebro, né bolsomínion?". Ivermectina é um medicamento utilizado para o tratamento de vermes. É sabido que os vermes se desenvolvem no intestino, em meio às fezes. A afirmação de Uo7 de que o dito remédio é para os vermes do cérebro de Uo3 equivale a dizer que Uo3 tem fezes no cérebro. Trata-se de uma afirmação violenta que desqualifica Uo3, ferindo-o moralmente, de forma insultuosa, expressando a aversão e o menosprezo pelo interlocutor. O vocativo aqui utilizado, "bolsominion", pejorativamente utilizado para caracterizar os que se posicionam favoráveis ao atual Presidente da República, defensor público dos remédios citados e fiador do discurso anti-vacina, irá suscitar a reação de Uo9, a qual comentaremos mais adiante, que evocará o neologismo "lularápio" em referência ao antigo Presidente, ideologicamente contrário ao atual.

Em resposta, Uo3, que defendia o uso de medicamentos contrapondo-se à vacinação, retoma a mesma estratégia violenta de Uo7, que o agrediu. Inicialmente, parece concordar com Uo7, ao dizer "Pode ser também!", mas, na sequência, agride Uo7, dizendo que "Já quem tem cérebro de jumento aí mesmo que não adianta!". A estratégia é diferente da do adversário, que se dirigiu diretamente a ele, com o emprego do possessivo "seu cérebro", direcionado à segunda pessoa do discurso (você). Uo3 deixa implícito a quem refere, utilizando o pronome relativo de caráter indefinido "quem" (aquele/aquela que). Como a resposta é dirigida explicitamente a Uo7, é possível inferir que "quem" refere a Uo7, no dizer de Uo3. Fechando o comentário, Uo3 expressa-se vitorioso na disputa de agressões, por meio de "Hehehe", vangloriando-se de um comentário bem-sucedido.

Uo8 chama a atenção de Uo3, também por meio de ironia - "que interessante!" - além do emprego de forma indefinida - "tem gente". Pela crítica à crença nos tratamentos medicamentosos contra a Covid-19 e com base no vocativo que nomeia Uo3, podemos afirmar que a crítica é dirigida a Uo3. Esse responde a Uo8, desqualificando seu comentário, sugerindo

que ele não está fundamentado em informações e notícia "de verdade". Chama assim Uo8 de mal informado e qualifica que a má informação de Uo8 o "faz passar vergonha", desqualificando a afirmação do usuário. Ao desqualificar o dizer de Uo8, Uo3 tenta desqualificar o próprio usuário, pois ao combater os argumentos do adversário, combate-se também a pessoa do adversário, afinal, se seus argumentos não são bons, ele também não é.

Encerra a discussão um comentário avaliativo da conversa como um todo. Nesse comentário, Ua9 usa a violência verbal, recorrendo ao mesmo campo semântico das violências empregadas nos comentários anteriores: asno, jumento, burro: "A jumentada do lularápio encantador de burros, tão que tão!!!". Nota-se aqui, além do neologismo referente ao ex-presidente que já apontamos, o caráter zoomórfico das ofensas, "jumentada" e "burros". O epíteto desqualifica tanto o ex-presidente quanto os seus seguidores, chamando-os não só de burros, mas também de ratos, ao intertextualizar o conto folclórico alemão "O Flautista de Hamelin" (CHAVES, 2012), por meio do qual faz uma analogia entre o político oposicionista e o flautista de Hamelin, personagem que encantou os ratos da cidade com sua flauta.

Observando o conjunto dos comentários, podemos afirmar que, de ambos os lados da polêmica, tanto a favor como contra vacinação/uso de medicamentos refutados pela ciência, a violência aparece da mesma forma, com o emprego de estratégias semelhantes. Na realidade, estabelece-se um campo semântico para a violência verbal, inaugurada por um dos usuários, o que permite inferir que o emprego de violência por um dos usuários estimulou o uso de violência verbal na continuidade do diálogo.

Considerações finais

No início deste artigo, estabelecemos como pergunta orientadora: a grande visibilidade das interações nas redes sociais altera o estatuto da violência nas situações de controvérsia argumentativa, como é o caso das polêmicas? Estabelecemos como objetivo para este trabalho investigar o estatuto da violência nas redes sociais, especificamente em interações marcadas pelo dissenso. Para tanto, analisamos uma breve interação polêmica na rede social *Facebook*, que, embora não tenha caráter generalizador, nos permite avançar algumas reflexões.

Depois de observar os poucos exemplos apresentados, de parte de uma interação em torno de um problema de sociedade, a importância da vacinação, acreditamos poder afirmar que a violência, no jogo polêmico, cumpre o papel sobretudo de construir ou reforçar uma imagem identitária para os grupos aos quais pertence cada usuário, no caso de nosso exemplo, negacionista/ pró-vacina; defensor da eficácia/não eficácia de determinados tratamentos, reforçando o caráter dicotômico da polêmica, que põe sempre dois polos de pontos de vista em interação. Observamos, ainda, que essa polêmica se desdobra para além do tema inicial e repousa sobre as questões políticas que estão na base das tomadas de posição dos indivíduos.

Com base nos exemplos observados, cremos poder afirmar que a violência fica à margem da discussão e argumentação central, no vacinar-se ou não se vacinar; eficácia/não

eficácia de determinados tratamentos, seu papel argumentativo é, na verdade mostrar pertencimento ao grupo (eu sou como vocês negacionista/pró-vacina). Mais do que cumprir a função da polêmica de persuadir um terceiro, conforme afirma Amossy (2017), tendo em vista que cada um está congelado no seu posicionamento, nas redes sociais, o uso da violência parece atuar para reforçar um posicionamento perante o grupo instituído do qual o usuário faz parte. É como se o usuário precisasse mostrar para os seus pares que ele atua efetivamente e, para tanto, agride o *adversário*. Essa agressão verbal está ligada ao *páthos,* que se manifesta, como vimos, nas escolhas linguísticas de palavras que agridem tanto os conhecimentos do outro quanto suas formas de ver o mundo. Essas escolhas, via de regra, apontam para uma zoomorfização do oponente ou de seu grupo, um dos ápices da agressividade, como se esse tipo de xingamento materializasse a punição física que se deseja impor ao adversário, cumprindo o papel de agredir, ainda que ninguém fique ferido fisicamente, daí a sua persistência. Assim, cremos poder afirmar que a violência, na polêmica analisada, cumpre o papel de marcar mais claramente a dicotomia entre os polos em discussão, adquirindo um estatuto de simulacro de uma guerra de fato que, em grande medida, mais contribui para ampliar a dissensão do que para estabelecer o diálogo na defesa de pontos de vistas diversos em busca de consenso.

Referências

AMOSSY, R. A Apologia da Polêmica. São Paulo: Contexto, 2017.

ANGENOT, M. *La parole pamphlétaire typologie des discour modernes.* Paris, Payot, 1982.

BOUSFIELD, D. Impolitenesse in the struggle for power. In: BOUSFIELD, D.; LOCHER, M. (Eds.). *Impoliteness in Language.* Berlin/NY: Mouton de Gruyter, 2008, p. 127-153.

CABRAL, A. L. T. Inteligência Retórica: violência e emoções na construção do ethos. *Verbum*, v. 9, n. 1, p. 49-64, 2020. Disponível em: https://revistas.pucsp.br/verbum/article/view/48365/pdf. Acesso em: 13 jun. 2021.

CABRAL, A. L. T. Violência verbal e argumentação nas redes sociais: comentários no *Facebook. Revista Calidoscópio*, v. 17, n. 3, p. 416-432, 2019. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/issue/view/789. Acesso em: 13 jun. 2021.

CABRAL, A. L. T.; MARQUESI, S. C.; SEARA, I. R. L'articulation entre le descriptif et les émotions dans l'argumentation en faveur de Dominique Strauss-Kahn. In: RABATEL, A.; MONTE, M.; RODRIGUES, M. G. S. (Eds.). *Comment les médias parlent des émotions l'Affaire Nafissatou Diallo contre Dominique Strauss-Kahn*. Limoges: Lambert-Lucas, 2015, p. 307-323.

CABRAL, A. L. T.; LIMA, N. V. Interações conflituosas e violência verbal nas redes sociais: polêmica em comentários no Facebook. *Revista (Con)textos Linguísticos*, v. 12, n. 22, Edição Especial Violência Verbal, p. 39-58, 2018. Disponível em: http://www.periodicos.ufes.br/contextoslinguisticos/article/view/20626/14231. Acesso em: 13 jun. 2021.

CABRAL, A. L. T.; LIMA, N. V. Argumentação e polêmica nas redes sociais: o papel de violência verbal. *Signo*. v. 42, n. 73, p. 86-97, 2017. Disponível em: http://dx.doi.org/10.17058/signo.v42i73.8004. Acesso em: 13 jun. 2021.

CABRAL, A. L. T.; ALBERT, S. A. de B. Quebra de polidez na interação: das redes sociais para os ambientes virtuais de aprendizagem. In: CABRAL, A. L. T.; SEARA, I. R.; GUARANHA, M. F. (Orgs.). *Descortesia e Cortesia: expressão de culturas*. São Paulo: Cortez, 2017, p. 267-294.

CHAVES, A. M. *O Flautista de Hamelin*. Tradução do poema de "The pied piper of Hamelin", de Robert Browning. *E-F@BULATIONS/E-F@BULAÇÕES*, v. 10, p. 54-72, 2012. Disponível em: https://ler.letras.up.pt/uploads/ficheiros/11443.pdf. Acesso em: 13 jun. 2021.

CNN BRASIL. Coronavac obtém registro definitivo na China e será usada na população em geral. *CNN Brasil*, São Paulo, 6 fev. 2021. Disponível em: https://www.cnnbrasil.com.br/saude/2021/02/06/coronavac-obtem-registro-definitivo-na-china-e-sera-usada-na-populacao-em-geral. Acesso em: 13 jun. 2021.

CULPEPER, J. *Impoliteness using language to cause offense*. Cambridge: Cambridge University Press, 2011.

CULPEPER, J. Reflections on impoliteness, relational work and power. In: BOUSFIELD, D.; LOCHER, M. (Eds.). *Impoliteness in Language.* Berlin/NY: Mouton de Gruyter, 2008, p. 17-44.

DANBLON, E. *La function persuasive* antropologie du discours rhétorique origines et actualité. Paris: Armand Colin, 2005.

DEVELOTTE, C. Décrire l'espace d'exposition discursive dans un campus numérique. *Le français dans le monde. Recherches et applications,* numéro spécial, p. 88–100, 2006. Disponível em: https://halshs.archives-ouvertes.fr/halshs-00151851. Acesso em: 13 jun. 2021.

GEORGES, F. *Identités Virtuelles les profils utilisateur du Web 2.0.* Mercuès: Editions Questions théoriques, 2010.

GOFFMAN, E. *Forms of talk.* Philadelphia: University of Pensylvania Press, [1974] 1981.

GRAHAM, S. L. ; HARDAKER, C. (Im)politeness in digital communication. In: CULPEPER, J.; HAUGH, M.; KADAR, D. Z. (Eds.). *The Palgrave Handbook of Linguistic (Im)politeness*. London : Palgrave Macmillan, 2017, p. 785-814.

HOUAISS, A.; VILLAR, M. S. *Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva, 2001.

KAUFMANN, J-C. *L'invention de soi:* une théorie de l'identité. Paris: Armand Colin, 2004.

KERBRAT-ORECCHIONI, C. Abordagem intercultural da polidez linguística: problemas teóricos e estudo de caso. In: CABRAL, A. L. T.; SEARA, I. R.; GUARANHA, M. F. (Orgs.). *Descortesia e Cortesia: expressão de culturas*. São Paulo: Cortez, 2017, p. 17-55.

KERBRAT-ORECCHIONI, C. Polidez e impolidez nos debates políticos televisivos: o caso dos debates entre dois turnos dos presidentes franceses. In: SEARA, I. R. (Ed.). *Cortesia: olhares e (re) invenções.* Lisboa: Chiado Editora, 2014, p. 47-82.

KERBRAT-ORECCHIONI, C. *Le discours en interaction.* Paris: Armand Colin, 2005.

LOCHER, M. Politeness research from past to future, with a special focus on the discoursive approach. In: AMAYA, L. F. ; HERNÁNDEZ-LÓPEZ, M. De la O. ; MORÓN, R. G. ; CRUZ, M. P. ; BORRERO, M. M. ; BARRANCA, M. R. (Eds.). *New perspectives on (Im)politeness and interpersonnal communication.* Newcastle : Cambridge Scholars Publishing, 2012, p. 36-60.

MAÍZ-ARÉVALO, C. Losing face on Facebook: linguistic strategies to repair face in a Spanish common interest group. In: BOU-FRANCHE, P.; BLITVICH, P. G-C. (Eds.). *Analysing Digital Discourse New Insights and Futures Directions.* Cham: Palgrave Macmillan, 2019, p. 283 – 309.

MEYER, M. Principia Rhetorica une théorie Générale de l'argumentation. Paris: Librairie Arthème Fayard, 2008.

MILLS, S. Sociocultural Approches to (Im)politeness. In *:* CULPEPER, J.; HAUGH, M.; KADAR, D. Z. (Eds.) *The Palgrave Handbook of Linguistic (Im)politeness*. London : Palgrave Macmillan, 2017, p. 41-60.

OLIVEIRA, A. L. A. M.; CABRAL, A. L. T. Batalhas de MC: um estudo sobre (im)polidez e categorização axiológica à luz da pragmática. *Revista de estudos da linguagem (UGMG)*, v. 28, n. 4, p. 1983-2004, 2020. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/16681. Acesso em: 13 jun. 2021.

PAVEAU, M-A. 2017. *L'Analyse du Discours Numérique*. Dictionnaire des formes et des pratiques. Paris: Hermann, 2017.

PLANTIN, C. *Dictionnaire de l'argumentation.* Une introduction aux études d'argumentation. Lyon: ENS Éditions, 2016.

SEARA, I. R.; CABRAL, A. L. T. O comentário elogiativo nas redes sociais: estratégias de cortesia valorizadora. *Revista da Associação Portuguesa de Linguística*, n. 3, p. 311-332, 2017. Disponível em: https://doi.org/10.26334/2183-9077/rapln3ano2017a17. Acesso em: 13 jun. 2021.

TECMUNDO. Brasil é o terceiro país com mais usuários no Facebook. *Tecmundo*, Curitiba, 27 fev. 2019. Disponível em: https://www.tecmundo.com.br/redes-sociais/139130-brasil-terceiro-pais-usuarios-facebook.htm Acesso em: 13 jun. 2021.

TERKOUFARI, M. Toward a unified theory of politeness, impoliteness and rudeness. In: BOUSFIELD, D.; LOCHER, M. A. (Eds.). *Impoliteness in Language*. Berlin, NY: Mouton de Gruyter, 2008, p. 45-74.

Recebido: 14/06/2021.

Aprovado: 28/07/2021.

Linha D'Água: São Paulo, v. 34, n. 02, p. 135-141, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p135-141

Resenha / Review

PAVEAU, Marie-Anne. *L'Analyse du Discours Numérique*. Dictionnaire des formes et des pratiques. Paris: Hermann Éditeurs, 2017, 400p.

Nathalia Akemi Sato Mitsunari*

nathalia.mitsunari@usp.br
https://orcid.org/0000-0003-1389-9337

A tecnologia e a informática desencadearam significativas mudanças nos modos de contato interpessoal, de trabalhar, de aprender e ensinar e de se informar. Segundo a professora Marie-Anne Paveau, nas últimas quatro décadas, houve a expansão de gigantes da web, como *Google, Microsoft, Apple, Facebook* e *Amazon*. Também as marcas são impressionantes, como a de 350.000 tuítes por minuto em dezembro de 2016 e a de 830.000 arquivos compartilhados por minuto pelo *Dropbox* nos Estados Unidos no mesmo ano. Poucos trabalhos, no entanto, foram desenvolvidos acerca do discurso digital. Muitos se limitaram a uma perspectiva logocêntrica e a uma representação antropocêntrica da máquina, isto é, buscando compreendê-la como suporte, sem considerar que o texto produzido a partir de ferramentas de *softwares* tem seu sentido coconstruído por determinações técnicas e algorítmicas.

Diante desse problema, Paveau elaborou um "Dicionário de Análise do Discurso Digital (DADN)"[1], publicado de dezembro de 2012 a julho de 2015 em *Technologies Discursives*[2], caderno de pesquisa. O objetivo foi fornecer um aparato teórico-metodológico para a análise do tecnodiscurso, demonstrando como as mudanças na mobilidade, nas formas de encontro e de se relacionar provocadas pela *Internet* afetam também o fazer científico, não só em sua maneira de produzir e em seu meio de circulação, mas também em seu trabalho teórico e metodológico. Nesse dicionário on-line, teve início a obra *Análise do Discurso Digital.* Dicionário das formas e das práticas.

* Doutoranda em Letras pelo Programa de Pós-Graduação em Filologia e Língua Portuguesa da Universidade de São Paulo – USP, São Paulo, Brasil.

1 Disponível em: https://technodiscours.hypotheses.org/category/dictionnaire-dadn Acesso em: 25 de mai. 2021.

2 Disponível em: https://technodiscours.hypotheses.org/ Acesso em: 25 de mai. 2021.

Esse importante dicionário acaba de ser traduzido para o português por Júlia Lourenço Costa e por Roberto Leiser Baronas e publicado pela editora Pontes. Essa publicação chega em hora muito oportuna, quatro anos depois do seu original em francês.

Está organizada em 31 verbetes, além da introdução: "Algoritmo", "Análise do Discurso Digital", "Ampliação", "Comentário", "Comunicação mediada por computador", "Compósito", "Corpus digital nativo", "Cor", "Ciberviolência discursiva", "Deslinearização", "Dualismo digital", "Ecologia do discurso", "Escrita digital", "Enunciador digital", "Ambiente", "Ética do discurso digital", "Extimidade", "Hashtag", "Hipertexto", "Imprevisibilidade", "Integridade contextual", "Leis do discurso digital", "Memória tecnodiscursiva", "Produso", "`Pseudonimato", "Relacionalidade", "Tecnodiscurso citado", "Tecnogênero de discurso", "Tecnografismo", "Tecnologia discursiva" e "Tuíte".

Cada um dos verbetes não só traz a definição de conceitos e a descrição de categorias de análise de tecnogêneros – uma tipologia, no caso dos verbetes "Comentário", "Ciberviolência", "Deslinearização", "Enunciador digital", "Tecnodiscurso citado", "Tecnogênero de discurso", "Tecnografismo" e "Tuíte" –, como também propõe um debate epistemológico. São citados estudos que tratam do discurso digital na França, nos Estados Unidos, na Inglaterra, na Itália, no Chile, na Coréia do Sul, na Índia, na Noruega, na Finlândia, em Portugal e no Brasil. São mencionadas três autoras brasileiras: Dóris de Arruda Carneiro da Cunha, Eni de Lourdes Puccinelli Orlandi e Cristiane Dias.

Paveau é membro da Pléaide[3], unidade de pesquisa pluridisciplinar que congrega geógrafos, linguistas e historiadores – da arte, do cinema, da literatura –, respondendo a diferentes vertentes das ciências humanas. Assim, encontramos em sua obra, conceitos como o de extimidade, da psicologia social de Serge Tisseron; o de produso, da teoria da comunicação de Axel Bruns; e o de ambiente, emprestado de teorias cognitivas do discurso. A autora assume, entretanto, a perspectiva da Análise Cognitiva do Discurso (PAVEAU, 2017, p. 167), apoiada em Lucy Suchman e em Edwin Hutchins. Compreende que o sistema cognitivo não é individual, porque se desenvolve a partir de um ambiente, com um conjunto de agentes humanos, sociais, e não humanos, instrumentos e objetos. Essa concepção de ambiente é central em seu estudo, porque se opõe à noção de condições de produção da Análise do Discurso francesa e a noções de contexto de teorias do texto e da interação, posicionando-se diante de uma longa tradição de estudos sobre a linguagem em uso, que se esforçou para delimitar um campo e uma unidade de análise para a interação verbal, considerando diferentes processos de inferência.

Paveau defende que essas teorias são dualistas, herdeiras de Saussure, porque distinguem o linguístico e o extralinguístico, o que as torna incapazes de compreender as especificidades dos discursos digitais nativos, cujo sentido é construído em um *continuum* entre linguagem e ambiente de produção[4]. Esse *continuum* está em sua forma compósita (sexto

[3] Disponível em: https://pleiade.univ-paris13.fr/ Acesso em: 8 de jun. 2021.

[4] Paveau menciona o trabalho de Dominique Cotte (2004 *apud* PAVEAU, 2017, p. 13), que se refere a co-enunciações tecnológicas.

verbete), que integra, ao mesmo tempo, linguagem e informática, visível nas *hashtags* e nos *hiperlinks*; imprevisibilidade (vigésimo verbete), pois sua forma compósita mista, em parte, produto de algoritmos ocultos ao enunciador, torna impossível a ele prever as formas de circulação de suas produções on-line; deslinearização (décimo verbete), pois não possui um eixo sintagmático específico, podendo se relacionar a inúmeros hipertextos, seja por meio de *hashtags*, seja por meio de *hiperlinks*; ampliação (terceiro verbete), que supera a razão gráfica, na medida em que as ferramentas da *web* permitem que sejam adicionados comentários e que sejam compartilhados os textos indefinidamente, de tal modo que a instância enunciativa do discurso digital nativo não tem limites demarcados; e relacionalidade (vigésimo sexto verbete), que diferente da noção de dialogismo, refere-se ao material e ao automático, decorrente da estrutura hipertextual da *web*.

Essas características colocam dúvidas quanto à investigabilidade do texto on-line ao linguista cujos princípios, métodos e objetos pertencem a contextos linguageiros pré-digitais. Parece-lhe difícil identificar dados observáveis de enunciador, enunciatário, tempo e espaço, porque a compreensão de um discurso digital nativo integra prolongamentos temáticos e metadiscursivos infindos, a partir de comentários (quarto verbete) de pseudônimos (vigésimo quinto verbete), quase sempre situados por dêiticos, como "ontem". Diante dessa dificuldade, Paveau constata duas tendências: ou se questiona a validade do texto digital, a partir de uma visão dualista (décimo primeiro verbete), para a qual há dois universos separados e de natureza distinta, um real e outro virtual, de simulações[5]; ou se avalia negativamente a produção on-line, julgando-a de baixa qualidade, denunciando-a como responsável pela perda da memória humana, pela diminuição da leitura, do contato com os saberes e com as formas de sociabilidade.

Para a autora, deve-se compreender que os modos de existência digitais e pré-digitais são diferentes, mas ambos os modos são reais e estão integrados, como estão outros modos de existência pertencentes a diferentes espaços sociais, nos quais evoluímos. Podem ser observadas, on-line e off-line, modificações em nossos discursos, em marcas linguísticas – como neologismos, neografismos e elipses – e em marcas extralinguísticas – como a rapidez, a ansiedade, a impulsividade e a agressividade (PAVEAU, 2017, p. 59-60). Além disso, não se pode entender a *Internet* como um terreno para coleta de dados de interesse de uma Linguística pré-digital. Deve-se analisar o discurso digital nativo como um objeto em si, considerando parâmetros tecnodiscursivos[6]. Daí a insistência no uso de termos compostos pelo prefixo *tecno-*, que evidencia que, constitutivamente, o texto on-line não é puramente linguagem; e a adoção do conceito de "produso" (vigésimo quarto verbete), que marca que, na *Internet,* não há fronteiras delimitadas entre produção e uso de textos. Os conteúdos são produzidos de forma colaborativa, por ampliação e relacionalidade.

---

[5] Esse dualismo se inscreve em uma longa tradição, que parte de Platão, passa por Descartes e se estende à atualidade, compreendendo o mundo em uma visão binária que distingue o material e o imaterial (PAVEAU, 2017, p. 122).

[6] Isabelle Pierozak (2014 *apud* PAVEAU, 2017, p. 10) é citada, para quem há uma diferença entre utilizar a *Internet* "for corpus", para a composição de um corpus, e "as corpus", como um corpus.

Em "Comentário", o foco são os parâmetros tecnodiscursivos para a análise do tecnogênero e os problemas da aplicação de categorias de análise pré-digitais no discurso digital nativo. O verbete é dividido em três partes: I. Descrição e definição. Um gênero renovado pela *web*; II. Tipologia dos comentários digitais; III. Questões de ética e de direito do discurso. Na primeira parte, o comentário é definido como uma das formas tecnodiscursivas mais frequentes e mais ricas da *Internet* – por isso, deve ser considerado um objeto central da Análise do Discurso Digital, para Paveau. Suas funções são múltiplas e evoluíram desde o século VI a.C., na Grécia, até os dias atuais. Continua como um lugar de exegese, de explicação, interpretação, sugestão e proposição, mas na *web,* transforma-se em formas inéditas em múltiplos planos, pela publicidade (estatuto técnico e jurídico), visibilidade (configuração tecnodiscursiva que define as relações entre os internautas e os enunciados), conversacionalidade e recursividade.

As teorias da interação definiram a conversação por um certo número de elementos e por marcadores conversacionais de abertura e de fechamento. No entanto, on-line, os segmentos de abertura de um comentário não correspondem a marcadores conversacionais, são janelas de comentários e metadados. Segmentos de fechamento, por sua vez, quase sempre não existem – enquanto as funções de resposta e de compartilhamento permanecem disponíveis, a conversação pode continuar indeterminadamente. É possível, ainda, comentar comentários ao infinito. Por conta dessa recursividade, para Paveau, é mais pertinente falar em conversacionalidade do que em conversação na Análise do Discurso Digital, e deve-se compreender que as formas de discurso citado na *Internet* são operações metadiscursivas de representação de atos de enunciação que comportam muito mais que formas de citação pela linguagem. A distinção enunciativa entre o discurso que cita e o discurso citado pode ser garantida apenas por dispositivos tecnológicos de compartilhamento e de comentário, ativado por botões que asseguram a função de representação dos atos de enunciação.

Nesse sentido, ainda na primeira parte do verbete, Paveau (2017, p. 38) identifica uma ausência de parâmetros digitais nas análises linguísticas feita por Malika Temmar (2013) e por Dóris Cunha (2014). No primeiro estudo, os autores de comentários são considerados enunciadores fictícios, por interagirem por meio de pseudônimos. Observa-se, assim, que a pesquisa foi feita a partir de uma posição exterior ao universo discursivo digital, em que a enunciação por meio de pseudônimo é regra. Examiná-lo como um apagamento enunciativo, além disso, é assumir uma concepção tradicional de autor, de patrimônio e assinatura, que limita a compreensão da responsabilidade enunciativa on-line. Já o segundo estudo, cujo objetivo era identificar as formas de diálogo com o outro, conclui que os comentários apresentam muito mais formas de alusão que de citação e que o discurso citado, nos *sites* da imprensa brasileira, é mostrado e direto. Desse modo, pouco se analisa a questão da ampliação enunciativa e discursiva na *web.*

Na segunda parte do verbete, são propostas quatro grandes categorias de comentários on-line, considerando particularidades tecnolinguageiras: 1. comentário relacional, em que se estabelece uma simples relação entre textos, do tipo fática; 2. comentário conversacional, em que se adiciona um conteúdo discursivo ou metadiscursivo a um texto primeiro; 3. comentário

deslocado, produzidos em espaços de mensagem privados; e 4. pseudocomentário, produzidos na ocasião do compartilhamento de textos e assim identificados porque não são reconhecidos pelos metadados como comentários. Na terceira parte do verbete, demonstra-se como o desconhecimento de particularidades tecnolinguageiras, como a distinção entre o público e o visível na *Internet,* ocasiona problemas de ética e de direito do discurso, na difusão de dados privados.

Constrói-se uma crítica aos princípios, métodos e objetos da Linguística contemporânea. Para Paveau, é essencial que os estudos discursivos contestem as concepções logocêntricas da Análise do Discurso tradicional e instaurem uma simetria – e não mais uma distinção – entre tecnolinguagem e realidade, inspirada na Antropologia Simétrica de Bruno Latour (1991 *apud* PAVEAU, 2017, p. 131). Para tanto, para Paveau, é preciso que os algoritmos sejam considerados, sendo incorporados à reflexão sobre o discurso digital nativo como *frames* ou representações – conceitos da Linguística Cognitiva. Eles produzem, pois, um certo número de regularidades e determinações que se aproximam, com ressalvas, das formações discursivas, definidas por Michel Foucault e por Michel Pêcheux (PAVEAU, 2017, p. 20). As ressalvas, na verdade, são muito importantes. Os *frames,* as formações discursivas em Foucault e as formações discursivas em Pêcheux implicam diferentes relações entre sujeito, mundo e linguagem.

De todo modo, destaca-se a expansão digital não como um novo canal de circulação nem como uma nova forma de codificação de conteúdo, mas como uma transformação ambiental, de agentes humanos e não humanos, que afeta e é afetada por estruturas sociais e suas relações. Defende-se uma perspectiva ecológica (décimo segundo verbete), que investigue os discursos, integrando linguagem, tecnologia, cultura, política e ética – proposta mais visível nos verbetes "Ciberviolência discursiva" (nono), "Enunciador digital" (décimo quarto), "Ética do discurso digital" (décimo), "Extimidade" (décimo sétimo verbete), "Pseudonimato" (décimo nono) e "Tuíte" (trigésimo primeiro). Nesses verbetes, demonstra-se como os valores mobilizados pelo discurso digital não são valores a ele inatos, são critérios morais de (in)aceitabilidade de representações que são atualizadas no espaço de ação e relação on-line.

Paveau (2017, p. 7) ressalta que a chamada revolução digital é, como a democracia ou a sexualidade, uma noção profundamente situada, sem quaisquer universalizações. Por isso, distingue quatro tipos de web: *web* 1.0, *web* implementada nos anos 1990, que se fundamenta em um sistema de distribuição de informações; *web* 2.0, *web* do começo dos anos 2000, que se apoia na interação de multiagentes, nas redes sociais e no compartilhamento multimidiático; *web* 3.0, *web* do início dos anos 2010, que organiza e faz uma curadoria de dados, graças ao desenvolvimento de metadados; e *web* 4.0, *web* que emerge nos anos 2020, integrando o on-line a elementos e atividades da vida cotidiana, a partir de aplicativos e *gadgets* (PAVEAU, 2017, p. 14-15).

Os exemplos, conceitos e categorias de análise apresentados ao longo do dicionário pertencem à web 2.0. Em relação à web 2.0, são identificadas quatro dificuldades para o

linguista: 1. para selecionar um *corpus* representativo, posto que, pela primeira vez na história da disciplina, pode-se acessar uma quantidade de enunciados tão grande; 2. para ter uma visão abrangente de hipertextos a partir de fragmentos parciais, uma vez que os dispositivos tecnodiscursivos não permitem chegar a um panorama global, tal qual se tem de um número de jornal impresso, por exemplo, e o leitor transforma o texto que lê, na medida em que escolhe um caminho de leitura, ao clicar em *hiperlinks*; 3. para considerar os efeitos da redocumentarização de textos e segmentos deles, que são compartilhados em massa em diferentes plataformas digitais, de tal maneira que, muitas vezes, a autoria se perde; 4. para definir o contexto tecnorelacional, que interfere no olhar do linguista, através de algoritmos, que fazem chegar a ele determinados textos – e não outros – e que destacam determinados enunciados, em detrimento de outros.

Perpassam essas dificuldades duas questões, interligadas e de extrema relevância, sobretudo, para o contexto pandêmico em que vivemos, no qual, certamente, a influência de todas as formas da *web* cresceu: a da "identidade calculada" (GEORGES, 2009 *apud* PAVEAU, 2017, p. 20) e a da memória tecnodiscursiva (vigésimo terceiro verbete). Sob a perspectiva ecológica do discurso digital nativo, as informações classificadas e hierarquizadas por algoritmos, a partir de traços de nossas atividades on-line, devem ter efeitos significativos em nossas vidas, fazendo previsões daquilo que pode nos interessar, para mostrá-lo de maneira priorizada. Assim, a tecnolinguagem prediz o porvir, nossos comportamentos, on-line e off-line, coconstruindo nossos enunciados e nossa identidade. Paveau sublinha que os algoritmos são elaborados pelo homem e, desse modo, a perspectiva pós-dualista da Análise do Discurso Digital não hipertrofia a ação da máquina, destituindo o homem de seu papel de sujeito. Insere, apenas, a tecnolinguagem nas questões relativas à linguagem e ao poder e à produção e reprodução do trabalho.

A memória discursiva, indubitavelmente, consolidou-se como objeto de estudo da Linguística pré-digital. Paveau menciona o estudo de Jean-Jacques Courtine, que elabora o conceito de memória discursiva como uma alternativa ao de formação discursiva, considerando a pluralidade de tempos históricos, passados, presentes e futuros, que se inscrevem em todo enunciado, que os faz circular. A memória discursiva digital, contudo, ainda não foi explorada com afinco, seja sob uma perspectiva discursiva, seja sob uma perspectiva textual. É objeto de reflexão apenas das ciências da informação e da comunicação.

Deve-se investigá-la, para a linguista francesa, considerando a estrutura hipertextual da *web*, os efeitos da redocumentarização de textos e a tensão entre a repetição e a rememoração. Essa tensão, descrita por Benoît Habert, se dá precisamente entre dois modos de construir a identidade individual e coletiva: a partir de uma compulsão em "mumificar" a vida diária, pelo *life logging* – expressão em inglês que designa o ato de registrar atividades diárias pessoais ou da vida pública, no *smartphone* ou no computador, utilizando *gadgets* e aplicativos dedicados a esse propósito; e a partir da reflexividade, capacidade do ecossistema digital de, por meio de algoritmos, redocumentalizar traços das atividades dos internautas.

A originalidade da obra, que poderia ter sido elaborada como um manual ou como um ensaio, como Paveau pensou em desenvolvê-la, está na exposição das problemáticas em torno do estudo do tecnodiscurso em Linguística por meio de verbetes, que precisam a perspectiva ecológica. O dicionário não tem um *corpus*, e não são feitas análises dos exemplos de discursos digitais nativos apresentados com as categorias descritas e discutidas. As soluções para essas problemáticas e os caminhos para persegui-las, muitas vezes, permanecem em aberto. O trabalho, assim, cumpre o papel de demonstrar a fecundidade heurística da Análise do Discurso Digital e de verificar algumas lacunas nas teorias apoiadas no discurso pré-digital, como um "manifesto programático", como o definiu Maingueneau[7].

Essas teorias devem considerar as reflexões desenvolvidas por Paveau acerca da tradição dualista e logocêntrica em que se inscrevem. Devem se atentar, apenas, a possíveis conflitos entre noções de sujeito, coletividade, identidade e suas relações com a linguagem e com o mundo, posto que a Análise do Discurso Digital de Maria-Anne e sua perspectiva ecológica estão fundamentadas em uma Análise Cognitiva do Discurso. No Brasil, o Laboratório de Estudos Urbanos da Universidade Estadual de Campinas (Labeurb – Unicamp) tem feito esse trabalho, em parceira com o Pléaide, no projeto internacional Rede franco-brasileira de Análise do Discurso Digital (A2DI)[8], coordenado pela Profa. Dra. Cristiane Dias e por Marie-Anne Paveau.

Recebido: 13/06/2021.
Aprovado: 09/07/2021.

[7] Disponível em: https://journals.openedition.org/aad/2554 Acesso em: 12 de jun. 2021.

[8] Disponível em: https://www.labeurb.unicamp.br/site/web/projeto/28 Acesso em: 08 de jun. 2021.