Prova de Habilitação: O Processamento da Língua Portuguesa em prol das Humanidades Digitais

Renata Vieira
CIDEHUS, Universidade de Évora

May 20, 2024

Resumo

Este documento, submetido à prova de Habilitação para o Exercício de Funções de Coordenação Científica, apresenta conjuntamente uma proposta de um programa de investigação (Parte1), e de um programa de pós-graduação (Parte 2). O documento elabora sobre o papel da língua portuguesa e suas tecnologias para o desenvolvimento das Humanidades Digitais. A proposta está contextualizado na investigação da proponente, em curso no CIDEHUS - Centro Interdisciplinar de História Culturas e Sociedades da Universidade de Évora. A última parte (Parte 3) apresenta as considerações finais.

Conteúdo

1 Programa de Investigação **4**

1.1 Introdução 4

1.1.1 Justificativa para o tema 4

1.1.2 Desafios 5

1.1.3 Organização da Parte 1 8

1.2 Preparação de bases textuais 8

1.2.1 Digitalização, Paleografia e Filologia 9

1.2.2 Normalização textual 9

1.2.3 Anotação 10

1.2.4 Adição de meta-dados 10

1.3 Transformação de textos em dados e geração de novos conhecimentos 11

1.3.1 Extração de informação 12

1.3.2 Representação de conhecimento e ontologias . . . 13

1.3.3 Dados ligados e dados *FAIR* 15

1.3.4 Tendências atuais em tecnologias da Linguagem . 15

1.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa no contexto do CIDEHUS 16

1.4.1 *Semantic alignment technologies for language understanding applied to digital humanities* 17
1.4.2 Memórias Paroquiais 17
1.4.3 *Monsoon*: O estado da Índia hispânico em perspetiva Digital (1580-1640) 18
1.4.4 História do Futuro 19
1.4.5 *Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices* . 19
1.4.6 *Intangible Cultural Heritage, Bridging the Past, Present, and Future* 20
1.5 Plano de investigação 20
1.5.1 Chronos, o Laboratório de Humanidades Digitais 22

2 Programa de pós-graduação 23
2.1 Eixos fundamentais 24
2.1.1 Interfaces: Humanidades e Tecnologias 24
2.1.2 Preparação de fontes textuais para Humanidades Digitais 24
2.1.3 Linguagem, Inteligência Artificial e Processamento da Língua Portuguesa 25
2.1.4 Projetos em Humanidades Digitais 26
2.2 Proposta de um Curso de Pós Graduação em Estudos da Linguagem e Humanidades Digitais 27
2.2.1 Áreas científicas 28
2.2.2 Objetivos 29

2.2.3 Competências a desenvolver 30
2.2.4 Destinatários 30
2.2.5 Metodologia de Ensino e de Avaliação 31
2.2.6 Unidades Curriculares 33
2.2.7 Dissertações e Teses 34

3 Considerações finais **35**

Parte 1

Programa de Investigação

1.1 Introdução

A parte 1 deste documento apresenta questões relevantes para a investigação relacionada ao Processamento de Linguagem Natural (PLN) na área de Humanidades Digitais (HD). Em particular, tem um enfoque no desenvolvimento de tecnologias para a língua portuguesa e suas aplicações. Está contextualizado em projetos atuais, em curso no CIDEHUS, Centro Interdisciplinar de História Culturas e Sociedades da Universidade de Évora, em que a proponente tem atuado de forma colaborativa com investigadores de áreas como Linguística, História, Arqueologia e Turismo.

1.1.1 Justificativa para o tema

A área de HD tem ganhado força e adeptos nas últimas décadas, em paralelo ao desenvolvimento de ferramentas digitais que ampliam as possibilidades de armazenamento, acesso e processamento de dados.

Essas capacidades estendem os horizontes de atuação de investigadores, permitindo a captura, organização e análise de um volume muito grande de informação. A partir dessa transição tecnológica, as Humanidades ganham visibilidade, atravessam fronteiras disciplinares e enfrentam desafios sem precedentes.

O CIDEHUS tem em seu plano estratégico a intenção de desenvolver essa área, e nos últimos anos organizou a criação do Laboratório de Humanidades Digitais[1] com o objetivo de impulsionar atividades nessa área.

A proposta aqui apresentada enfatiza o tratamento textual com base em tecnologia da linguagem, área em que vários projetos do CIDEHUS têm se destacado (como será visto nas próximas secções). As bases textuais abordadas nesses projetos possuem relevância por suas características históricas, literárias, e sociais.

1.1.2 Desafios

Projetos na área de HD, baseados em fontes textuais, apresentam variações nos períodos históricos, nos suportes (manuscritos em papel, impressos, fotografados, etc), bem como nos estágios de digitalização. A digitalização pode variar entre imagens digitais, textos em PDF e textos digitalizados em outros formatos. Todas essas questões adicionam esforços extras ao processamento textual. Desta forma, a área requer preparação e organização das fontes, que, só depois de transcritas e digitalizadas, podem ser submetidas a processamentos mais avançados.

[1] https://sites.google.com/view/hdlabcidehus

O CIDEHUS possui vários projetos em curso que enfrentam diferentes etapas desse processo. A figura 1.1 ilustra a variedade de fontes em estudo e sua relação com outros temas como história do comércio na Ásia, filosofia, geografia e arqueologia.

Figura 1.1: Exemplos de bases textuais, objetos de estudo por investigadores do CIDEHUS que empregam variadas tecnologias de linguagem em seus processos. Fonte: Vieira et al, apresentação realizada em *Workshop on Digital Humanities and NlP - PROPOR*, 2022.

Na figura 1.2, o processo de adaptação de fontes textuais históricas no percurso ao longo do tempo é ilustrado, com base na trajetória das Memórias Paroquiais. Nesse exemplo temos uma fonte original manuscrita (1758), que foi transcrita para impressão (por exemplo, em 1943), impressa com comentários adicionados (2002), transcrita para uma versão digital (2016), utilizada para geração de listas de entidades nomeadas (2021), com possibilidades adicionais de processamento previstas para o futuro.

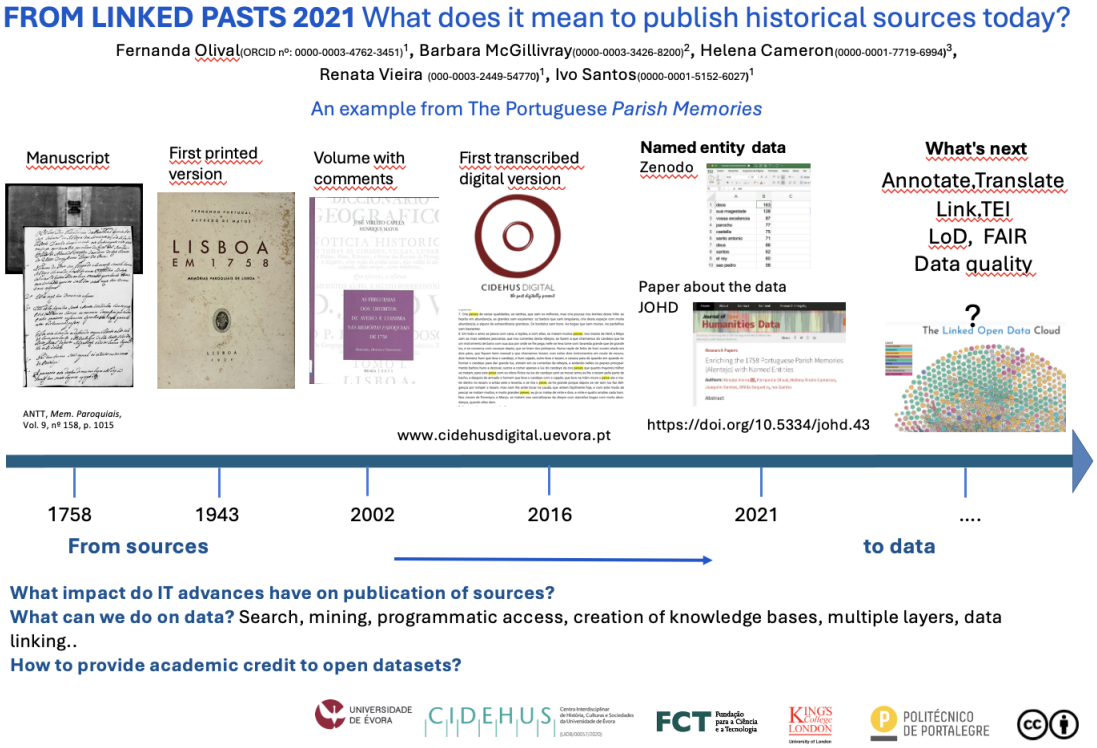

Figura 1.2: Exemplos do percurso das Memórias Paroquiais, atualizada ao longo do tempo, e que em suas fases mais atuais, conta com recursos de tecnologias de linguagem para extração de informação. Fonte: Olival et al, Apresentação realizada no evento Linked Pasts, 2021.

Nessa fonte, um trabalho extenso foi realizado com o original ao longo dos anos até estar apto ao processamento de linguagem natural, para assim poder obter-se dados estruturados dos textos, disponibilizar e associar os dados encontrados com outras bases. Nesse percurso, a semântica dos dados é um aspeto fundamental. Para enfrentar os desafios desse processo, vários conhecimentos são necessários, e a área requer uma abordagem inter ou transdisciplinar.

1.1.3 Organização da Parte 1

A parte inicial deste documento destina-se à apresentação do programa de investigação, e está organizado como segue. A Secção 1.2 aborda os tópicos e desafios relacionados à preparação das fontes, desde a digitalização até a anotação. A Secção 1.3, discute os processos de transformação de fontes em dados, e a necessidade de organização semântica dos dados. A Secção 1.3.4 comenta sobre as tendências atuais em PLN, uma vez que a área passa por uma revolução tecnológica importante, aspetos que não podem ser desconsiderados aqui. A Secção 1.4, lista alguns projetos relacionados a HD, em curso no CIDEHUS. A Secção 1.5 apresenta o plano de investigação, centrado na condução de projetos de investigação financiados e em curso no CIDEHUS.

1.2 Preparação de bases textuais

Na área de PLN, geralmente se parte de bases textuais digitais, que representam a construção de conhecimento mais atual onde versões digitalizadas são disponíveis, em versões *online* de jornais, de revistas académicas ou mesmo em redes sociais. Porém, os trabalhos em HD, que se baseiam em textos, podem requerer uma fase de preparação intensiva, quando os dados não se encontram em um formato ou *media* processável por computador. Essa preparação pode se beneficiar das atuais técnicas de Inteligência Artificial (IA) e PLN para um melhor acesso às fontes. A preparação pode requerer uma variedade de processos no tratamento das fontes e sua informação:

1.2.1 Digitalização, Paleografia e Filologia

A digitalização de documentos, que geralmente emprega técnicas de OCR (*Optical character recognition*), ou HDR (*High-dynamic-range imaging*), é o passo inicial para preparar uma base de estudo para o tratamento computacional, e assim possibilitar a sua leitura por meio de programação [Mittal and Garg, 2020, Van Strien et al., 2020]. O processo pode consumir uma fatia de tempo considerável dos projetos, uma vez que podem ser necessárias revisões e correções nos processos.

A natureza dos originais pode requerer especialidades nas áreas de Paleografia e Filologia, [Ferreira et al., 2024], [Castillo Gómez, 2024], [Gonçalves and Banza, 2013], que abordam sistemas de escrita histórica, a historicidade de manuscritos, e sua decifração, a linguagem em fontes históricas orais e escritas, as interações entre história e linguística.

1.2.2 Normalização textual

A normalização para uma língua padrão e contemporânea pode ser necessária em situações em que fontes históricas são estudadas, ou em textos atuais, utilizados em redes sociais, que usam muitas abreviações, símbolos associados à emoções, ou comentários como *hashtags*. Apesar de ser um problema de base, a ser tratado ainda em fase de preparação, não há soluções computacionais prontas, capazes de lidar com qualquer tipo de texto que necessite esse processo [Cameron et al., 2023].

1.2.3 Anotação

Muitas vezes o estudo de uma fonte ou *corpus* necessita a adição de informações extras sobre o registo, os processos de inclusão dessas informações adicionais são chamados de anotação. Os processos podem ser automáticos ou manuais. No entanto, geralmente o desenvolvimento de um processo automático requer um processo de anotação manual anterior, quer para o seu desenvolvimento (treino de algoritmos de aprendizado) ou para a avaliação da correção da anotação realizada por máquina. Podem ser adicionadas informações de análise linguística: morfológica, sintática ou semântica (por exemplo, identificando entidades ou eventos) [Freitas, 2024].

1.2.4 Adição de meta-dados

Meta-dados são os meios pelos quais as informações adicionais são armazenadas, podem ser informação extra-textuais importantes para a estruturação, armazenamento dos textos. Os meta-dados podem trazer informações de contexto, como autores, datas, locais de produção, origem, ou ainda informações como volume, páginas, etc. Servem também para registar interpretações sobre o conteúdo de uma coleção de interesse provenientes do processo de anotação descrito acima. Há diferentes propostas de formatos para incorporá-los aos textos [Nair and Jeeven, 2004].

1.3 Transformação de textos em dados e geração de novos conhecimentos

Os processos de transformação de textos em dados organizados possibilitam ao investigador realizar análises diferenciadas sobre textos, por exemplo, ao agrupar tipos específicos de informação, ou rapidamente quantificar fenómenos observáveis.

Assim, com a evolução dos estudos em HD, são produzidos dados mais elaborados, mais numerosos, e úteis aos investigadores em humanidades. Amplia-se a capacidade de análise, pela possibilidade de trabalhar com um grande volume de informação e pela capacidade de organizar essa informação, de forma mais ágil e rápida, em estruturas bem definidas.

O PLN é a área responsável por possibilitar essa transformação de textos em dados. São comuns nesses processos a presença dos seguintes elementos:

- extração de informação para a criação de conjunto de dados,
- uso de técnicas de representação e organização de conhecimento, para organizar os dados,
- criação de bases de dados interligadas, de forma a estimular o reuso e a colaboração.

Contudo, tais processos, quando aplicados em pesquisas na área das humanidades, requerem não apenas ferramentas atuais de PLN,

mas também uma interação mais próxima com os investigadores das respetivas áreas. Isso se faz necessário para o desenvolvimento das adaptações de ferramentas a diferentes objetivos e para a construção de interfaces adequadas ao uso das mesmas. Essa construção requer uma elaboração interdisciplinar. É importante aliar os interesses dos investigadores em humanidades às possibilidades mais atuais e eficientes de obtenção de informação por meio da aplicação de tecnologias da linguagem, especialidades dos investigadores em PLN. Discutiremos a seguir os elementos, mencionados acima, respeitantes à transformação de textos em dados.

1.3.1 Extração de informação

Com textos das áreas das humanidades que estejam já digitalizados e normalizados, podem ser aplicadas ferramentas de PLN já desenvolvidas para diferentes tarefas como, por exemplo, reconhecimento de entidades nomeadas, extração de eventos [Claro et al., 2024], resolução de correferências [Fonseca et al., 2024], e sistemas de perguntas e respostas [Cortes et al., 2024].

A extração de informação é o processo mais fortemente relacionado ao objetivo de transformação de dados não estruturados (textuais) em dados estruturados (tabelas ou banco de dados). Enquadram-se nela as tarefas de reconhecimento de entidades nomeadas e identificação e classificação de eventos, aplicadas, por exemplo, nos projetos descritos na Secção 1.4.

O reconhecimento de entidades nomeadas possibilita a identificação de atores importantes e de outras personagens de um dado período; possibilita também mapear áreas de atuação, fazer relações com questões de cartografia e sistemas de informação geográfica. As instituições, de diferentes perfis, também podem ser identificadas. Se usarmos técnicas mais avançadas, os relacionamentos entre essas entidades também poderão trazer informações relevantes para os investigadores. A identificação de eventos e respetiva classificação podem ajudar a abordar zonas e contextos significativos, como, por exemplo, no projeto *Monsoon* (Secção 1.4) que realiza uma análise dos eventos inerentes a conflitos na coleção "Livro das Monções".

1.3.2 Representação de conhecimento e ontologias

Ontologias são artefactos conceituais interpretados com diferentes nuances em filosofia, lógica e sistemas de informação. A ideia central é representar conhecimento de forma estruturada. Em HD, a nuance mais comummente relacionada é aquela empregada em sistemas de informação, ontologias como estruturas conceituais que podem ajudar na organização dos dados e na comunicação entre seres humanos e máquinas [Guarino, 1998]. As ontologias são essenciais para a organização semântica dos dados.

O nível de detalhe de uma ontologia pode variar entre estruturas taxonómicas, definição de instâncias, relações e axiomas lógicos. Elas auxiliam a esclarecer (e combinar) as diferentes concetualizações que

compõem diferentes disciplinas.

Ontologias têm sido usadas em diversos domínios do conhecimento em diferentes tarefas: organização e anotação de grandes quantidades de dados (anotação semântica); integração de dados de diversas fontes (integração semântica), representar conhecimento de domínios complexos, dar ancoragem para o raciocínio automático, e suportar buscas em grandes quantidades de dados (busca semântica).

Em HD, além das tarefas citadas acima, ontologias têm sido utilizadas para para representação de meta-dados, para melhor especificar as informações codificadas em cada *corpus* ou fonte. Essas descrições formais permitem a integração de dados, de maneira independente de software e esquema. Um passo essencial para melhorar a qualidade dos dados é usar vocabulários padronizados e ontologias para representação de dados e meta-dados [Guizzardi, 2020].

Uma ontologia bastante referenciada em HD, em especial na área do património cultural é o CIDOC-CRM[2]. É uma ferramenta teórica e prática para a integração de informação. O objetivo é ajudar os investigadores, os administradores e o público a explorar questões complexas, relacionadas ao passado, disponíveis em conjuntos de dados diversos e dispersos. O CIDOC-CRM fornece definições e uma estrutura formal para descrever os conceitos e relações implícitos e explícitos, utilizados em documentações do património cultural.

[2]https://www.cidoc-crm.org

1.3.3 Dados ligados e dados *FAIR*

Os esforços de partilha de dados produzidos nas pesquisas ainda têm um longo caminho a percorrer em termos de padronização. É muito importante tornar os dados compatíveis com os princípios *FAIR*

- *Findability*/Encontrabilidade,
- *Accessibility*/Acessibilidade,
- *Interoperability*/Interoperabilidade,
- *Reuse*/Reutilização).

Esses princípios [Wilkinson et al., 2016] correspondem a um conjunto de 15 recomendações que visam facilitar a reutilização de dados por humanos e máquinas.

1.3.4 Tendências atuais em tecnologias da Linguagem

Nos últimos cinco anos, aproximadamente, a área de PLN passou por importantes evoluções tecnológicas, com a construção dos grandes modelos de linguagem (*large language models*), os LLMs [Paes et al., 2024]. Esses modelos são gerados por processamento intensivo de grandes quantidades de texto, sendo capazes de capturar os padrões gerais de uma linguagem, ou conjunto de linguagens, pois há modelos multilingues. Os modelos possuem a capacidade de gerar textos bem organizados a partir de instruções. Embora de limitações conhecidas, como a baixa aderência aos fatos (limitação batizada de alucinação),

foi notório o impacto do ChatGPT, da OpenAI[3], em muitas áreas de atuação. Os grandes modelos de linguagem semelhantes ao que se baseia o ChatGPT se proliferam e tem sido assunto dominante em muitas conferências científicas, não só na área de computação.

Grandes empresas e grupos académicos desenvolvem modelos alternativos. Em geral são recursos de alto consumo computacional.

Uma família desses modelos originou-se a partir do modelo denominado BERT, e para o português vimos aparecer os modelos BERTimbau [Souza et al., 2024], Albertina PT [Rodrigues et al., 2023], GlórIA [Lopes et al., 2024].

Atualmente o desenvolvimento de recursos para o processamento da língua, baseados em aprendizado a partir de dados, incorporam algum desses modelos. Em alguns dos nossos trabalhos mais recentes algumas alternativas desses modelos foram avaliadas [Santos et al., 2024] para a tarefa de reconhecimento de entidades nomeadas.

A seguir serão apresentados alguns projetos em curso que seguem os processos e utilizam tecnologias e ferramentas mencionados acima.

1.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa no contexto do CIDEHUS

Pode-se considerar que o programa em investigação proposto já encontra-se em desenvolvimento. Foi sendo construído nos 4 anos de atuação da

[3]https://chat.openai.com

proponente no CIDEHUS, em que procurou conhecer os projetos em curso, contribuir com elementos de PLN e integrá-los em uma visão conjunta. A seguir, são apresentados os trabalhos envolvendo HD e PLN com foco em língua portuguesa, desenvolvidos no contexto do Laboratório Chronos, de Humanidades Digitais[4]. Os dois últimos projetos são multilingues e consideram o português entre outras línguas.

1.4.1 *Semantic alignment technologies for language understanding applied to digital humanities*

Esse projeto, com financiamento FCT CEECIND/01997/2017, pode ser considerado o projeto inicial a partir do qual se desenvolveram as relações com os outros projetos do CIDEHUS.

O projeto aborda questões de alinhamento semântico para composição de fontes de conhecimento heterogéneas (dicionários, ontologias, bases textuais, bases de dados) com o objetivo de facilitar o acesso ao conhecimento especializado.

1.4.2 Memórias Paroquiais

As Memórias Paroquiais constituem uma coleção muito estudada em Portugal. A versão digitalizada dos microfilmes dos originais está disponível no Arquivo Nacional da Torre do Tombo. Existem diversos livros impressos reproduzindo partes da coleção. Uma versão digital parcial foi disponibilizada, gratuitamente, através do Portal Digital do

[4]https://sites.google.com/view/hdlabcidehus

CIDEHUS[5]. Isso possibilitou a aplicação de técnicas de PLN no contexto do projeto *CIDEHUSDigital: criar e desenvolver uma equipa de anotação*[6], coordenado pela Profa. Fernanda Olival. Um conjunto de dados foi construído automaticamente usando sistemas previamente desenvolvidos para reconhecimento de entidades nomeadas [Santos et al., 2019]. Essa base e seu processo de construção estão descritos em [Vieira et al., 2021]. Posteriormente, essas anotações iniciais foram aperfeiçoadas e serviram de exemplos para construção de um modelo de anotação automática [Santos et al., 2024], com uma maior precisão. O projeto conta com financiamento do CIDEHUS.

1.4.3 *Monsoon*: O estado da Índia hispânico em perspetiva Digital (1580-1640)

Extração de eventos é outra tarefa de PLN com recursos desenvolvidos para o português [Sacramento and Souza, 2021]. O desafio é adaptar essas ferramentas à língua portuguesa do século XVII, esforço que está sendo explorado no projeto *Monsoon: o Estado da Índia hispânico em perspetiva digital (1580-1640)*[7], coordenado pela Profa. Ana Sofia Ribeiro e financiado pela FCT como projeto exploratório. Através do sistema de extração de eventos, temos acesso a estatísticas de classificação que nos permitem uma perceção mais evidente dos assuntos representados na coleção [Albuquerque et al., 2024]. Esse projeto utiliza

[5]http://www.cidehusdigital.uevora.pt
[6]https://sites.google.com/view/cdnota
[7]https://sites.google.com/view/monsoonpt

também a anotação de entidades nomeadas de quatro tipos: pessoas, locais, grupos de pessoas e instituições.

1.4.4 História do Futuro

No projeto em curso sobre a *História do Futuro*[8] [Banza, 2022], de Padre António Vieira, coordenado pela Profa. Ana Paula Banza, as etapas de transcrição, anotação e extração são aplicadas com interesse em exploração linguística do corpus e para o estudo das alternativas de sua composição. O projeto contou com financiamento do CIDEHUS.

1.4.5 *Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices*

Em projeto recente de colaboração internacional, lidamos com o estudo do problema de desinformação. É o projeto HYBRIDS *Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices*[9], uma rede *Marie Curie* coordenada pelo Prof. Pablo Gamallo em Santiago de Compostela, com financiamento da Comissão Europeia, HORIZON-MSCA-2021-DN-01. No contexto desse projeto investiga-se o papel da tecnologia no combate a desinformação, em temas críticos como imigração [Marino et al., 2024], clima e saúde.

[8] https://sites.google.com/uevora.pt/hdof
[9] https://hybridsproject.eu

1.4.6 *Intangible Cultural Heritage, Bridging the Past, Present, and Future*

O estudo do patrimônio cultural é tema do projeto INT-ACT, *Intangible Cultural Heritage, Bridging the Past, Present, and Future*[10] coordenado pelo Prof. Masood Masoodian da Universidade AALTO na Finlândia, financiado pela Comissão Europeia, HORIZON-CL2 2023-HERITAGE-01-04. Projeto iniciado em janeiro de 2024. Trata da comprensão das possíveis interpretações do patrimônio cultural e sua incorporação em produtos digitais de divulgação do patrimônio, elaborando nas dimensões do ambiente, das experiências e das emoções.

1.5 Plano de investigação

O plano de investigação está baseado na continuidade de projetos em andamento. Os projetos têm um núcleo comum, lidam com o processamento da língua, a organização semântica do conhecimento refletido em bases textuais e sua análise auxiliada por tecnologias de linguagem, cada um com diferentes temáticas. Envolvem preparação, normalização e anotação de textos, extração de informação, e construção de ontologias com apoio de técnicas de IA e PLN. São desenvolvidos de forma colaborativa com participação de alunos de doutorado, mestrado e bolseiros de investigação.

Pode-se organizar os projetos em:

[10] https://int-act.aalto.fi/index.html

- Projeto nuclear 1) *Semantic alignment technologies for language understanding applied to digital humanities* no qual sou IR; Nessa investigação procura-se compreender os problemas comuns no tratamento de fontes para seu processamento, que envolvem obter a fonte, digitalizar, adaptar, processar, analisar resultados e produzir novos conhecimentos. De acordo com projetos derivados (listados a seguir) atende-se aos objetivos de conhecer a história, interpretar a literatura, valorizar o património, combater a desinformação.

- Projetos Nacionais 2)*Monsoon* (Secção 1.4.3), 3) *CD-Nota* (Secção 1.4.2), onde atuo como co-IR e investigadora respetivamente;

- Projetos Europeus 4) Hybrids - *Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices*, no qual sou IR na Universidade de Évora, líder do *Work Package 1: Hybrid Intelligence*, orientadora principal do doutorando Erik Marino e co-orientadora do doutorando Davide Bassi da Universidade de Santiago de Compostela; 5) INT-ACT, *Intangible Cultural Heritage, Bridging the Past, Present, and Future*, no qual sou IR na Universidade de Évora e líder do *Work Package 1: Formalised Knowledge*, orientadora do bolseiro Antonio Diniz e co-orientadora da bolseira Camila Campos.

Além da execução dos projetos listados acima, situa-se no plano de investigação a co-orientação do doutorando Ivo Santos em seu projeto na área de Extração de Informação em relatórios de arqueologia, finan-

ciado pela FCT, e em linha com as questões do processamento da língua portuguesa [Santos and Vieira, 2021], a co-orientação do mestrando em história, Rafael Prezado, e a co-coordenação do Laboratório Chronos, de Humanidades Digitais.

1.5.1 Chronos, o Laboratório de Humanidades Digitais

O grupo de investigadores interessados na área de Humanidades Digitais do CIDEHUS, a seguir de um encontro online realizado em 2022 [Vieira and Banza, 2022], organizou-se para a constituição do laboratório Chronos. As atividades desse grupo estão brevemente refletidas na sua página web[11], que também apresenta mais detalhes sobre os projetos aqui mencionados. Dessa forma, constitui parte do plano de trabalho coordenar e dinamizar as atividades do laboratório.

[11] https://sites.google.com/view/hdlabcidehus

Parte 2

Programa de pós-graduação

Nesta secção, é apresentada uma estrutura conceptual inicial e geral para um programa de pós-graduação, interdisciplinar, centrado no estudo da língua e suas tecnologias para o desenvolvimento de projetos em humanidades digitais.

A proposta está organizada em quatro pilares, apresentados a seguir. Os 3 primeiros eixos são gerais, referem-se a entender o trabalho interdisciplinar, as características principais das relações entre as Humanidades e as areas tecnológicas, os estudos da linguagem, as características das fontes textuais, e o seu tratamento, e o processamento da língua. O quarto eixo visa reconhecer esses aspetos em aplicações, a partir de experiência em projetos.

2.1 Eixos fundamentais

2.1.1 Interfaces: Humanidades e Tecnologias

A construção conjunta estabelecida por meio de diálogo entre áreas complementares é crucial. Importa ter conhecimento dos problemas a serem tratados, assim como das técnicas possíveis de serem empregadas, o que requer interagir com diferentes especialistas. Muitas vezes é possível aplicar soluções já desenvolvidas, em outras situações poderá ser necessário desenvolver novas soluções para problemas específicos. Para esse eixo temos como áreas envolvidas: linguística, história, sociologia e informática, com as seguintes componentes:

- **Introdução a Humanidades Digitais:** estudar as premissas da área, suas principais aplicações e também as visões críticas;
- **Multi, inter e transdisciplinaridade:** estudar as visões de cada uma dessas dimensões e preparar para embarcar em projetos multi, inter e transdisciplinares;
- **Evolução da comunicação homem-máquina:** estudar a evolução digital, a revolução digital e os seus impactos na sociedade.

2.1.2 Preparação de fontes textuais para Humanidades Digitais

Conhecer a variedade de fontes, suas necessidades de adaptação para o contexto atual e as diferentes possibilidades de uso de tecnologia para

aperfeiçoar essa adaptação, devem ser contempladas. Soluções manuais, computacionais ou mistas precisam ser compreendidas. Para esse eixo, as áreas envolvidas são: linguística e história, e os componentes:

- **Paleografia:** estudar sistemas de escrita histórica, a historicidade de manuscritos, e sua decifração, incluindo a análise de caligrafia de forma manual e digital;
- **Filologia:** estudar a linguagem em fontes históricas orais e escritas, compreender as interações entre história e linguística;
- **Digitalização textual:** conhecer as ferramentas de digitalização e os problemas e alternativas de construção de bases textuais digitais.

2.1.3 Linguagem, Inteligência Artificial e Processamento da Língua Portuguesa

Aqui a linguagem será estudada sob diferentes ângulos, estruturais, aplicados, e sociais. Uma importante tecnologia por traz de muitos dos avanços computacionais é a Inteligência Artificial e os seu sub-domínios Processamento de Linguagem Natural e de Imagem. Compreender de forma conceptual e saber colocar em prática alguns recursos é o objetivo desse eixo, que tem como áreas: informática e linguística, e como componentes:

- **Estudos da Linguagem:** estudar as estruturas presentes no estudo da língua, fonética, morfologia, sintaxe, semântica, pragmática,

bem como a relação do estudo da linguagem com outras áreas psicologia, sociologia, e computação;

- **Introdução à Inteligência Artificial, PLN e Modelos de linguagem:** : estudar a área de inteligência artificial, suas subdivisões e abordagens, apresentar conceitos básicos de programação para as Humanidades, entender a área de PLN, em especial o processamento da língua portuguesa, e os princípios das novas tecnologias baseadas em modelos de linguagem;
- **Dados abertos, dados ligados, dados FAIR e ontologias:** entender a importância da ciência aberta, estudar e aplicar as boas práticas relacionadas a disponibilização de dados de investigação e o uso de meta-dados.

2.1.4 Projetos em Humanidades Digitais

Nesse eixo propõe-se uma componente baseada em desenvolvimento de projetos. Conhecer os projetos de HD existentes, em especial aqueles em curso na Universidade, integrar ou definir um sub-projeto ou projeto original. Criar um plano de projeto, a ser aprovado, executar e apresentar os resultados. Áreas envolvidas: humanidades em geral, informática.

2.2 Proposta de um Curso de Pós Graduação em Estudos da Linguagem e Humanidades Digitais

Com base nos pilares apresentados na Secção 2.1, propõem-se um curso de Pós-Graduação em Estudos da Linguagem em Humanidades Digitais, de formação transdisciplinar: humana, científica e tecnológica, centrado na área da linguagem e suas tecnologias, com aplicação no desenvolvimento das Humanidades Digitais.

Poderá atender tanto os licenciados em cursos de Humanidades que querem se desenvolver com base nas novas tecnologias de linguagem, como os de áreas tecnológicas que querem ampliar seu conhecimento em Humanidades e desenvolver novas aplicações. Com a crescente influência da Inteligência Artificial na sociedade o curso visa preparar aqueles que procuram uma formação integral, mista, técnica e humana.

O plano curricular deverá considerar a articulação interdisciplinar entre as componentes teóricas, e a articulação transdisciplinar em metodologia e prática, relevantes ao desenvolvimento das novas tecnologias de linguagem e das Humanidades Digitais.

A ideia é proporcionar a discussão sobre projetos em desenvolvimento na área de Humanidades Digitais, com exemplos de projetos financiados, nacionais e internacionais, bem como projetos em curso na Universidade de Évora. Particular atenção será dada a projetos baseados em fontes textuais. Há possibilidade de abranger uma variedade de elementos das Humanidades, projetos baseados em fontes históricas,

bases literárias, mas também de cunho social (por exemplo, baseados em análise de redes sociais), património, e turismo. No estágio atual do desenvolvimento de tecnologias da linguagem, é particularmente importante identificar projetos que aplicam essas tecnologias e que estejam em sintonia com questões de divulgação de resultado baseado em dados abertos, bem como questões éticas.

Os estudantes poderão adquirir competências de análise, crítica, e conhecer as fases de desenvolvimento de um projeto transdisciplinar, desafios para as fases de proposta, gestão, execução, elaboração de relatórios, e saber reconhecer as boas práticas em disponibilização de dados. Irá se desenvolver uma abordagem teórica-prática e crítica dos conhecimentos nas temáticas das Humanidades Digitais, com a especificidade de serem baseados em análise textual.

2.2.1 Áreas científicas

Propõem-se um curso com abordagens inter e transdisciplinares, sendo as áreas nucleares Linguística, Informática, História e Sociologia. Este é, portanto, um projeto apenas inicialmente idealizado, pois na presente proposta está apresentado sob a ótica de um único investigador, seu desenvolvimento concreto deverá incorporar visões das demais áreas.

Conhecer a relevância dos estudos da linguagem em projetos em Humanidades Digitais
Saber identificar potenciais de inovação nas Humanidades
Conceber e executar projetos na área de Humanidades Digitais
Conhecer o valor do conhecimento sobre a linguagem no mercado das tecnologias da língua
Ter experiência de transdisciplinaridade

Table 2.1: Objetivos do ponto de vista dos conhecimentos

Desenvolver a capacidade de elaboração de projetos transdisciplinares
Desenvolver a habilidade para comunicar resultados de projetos e sua relevância
Desenvolver aptidão para organização e planificação do trabalho transdisciplinar
Estimular o sentido de ética perante o trabalho envolvendo novas tecnologias
Desenvolver a compreensão e o interesse por uma nova maneira de construção de conhecimento
Desenvolver as capacidades de interação crítica com a tecnologia

Table 2.2: Objetivos do ponto de vista da aquisição das habilidades

2.2.2 Objetivos

Nas tabela 2.1 e 2.2 são listados alguns objetivos em linha com essa proposta. De forma geral foca-se em desenvolver uma base sólida sobre a linguagem e compreender a relevância das tecnologias atuais para as áreas das Humanidades em um contexto transdisciplinar.

2.2.3 Competências a desenvolver

Nesse programa, poderão se desenvolver profissionais e académicos que possam lidar com o conhecimento registados em textos, nas suas diversas formas contemporâneas e arcaicas, com o apoio de tecnologia. O curso virá ajudar a compreender a diversidade e complexidade dos estudos da linguagem, do desenvolvimento da Inteligência Artificial e e do Processamento de Linguagem Natural e sua relevância no desenvolvimento das Humanidades; reconhecer a importância de promover a interdisciplinaridade e a transdisciplinaridade; compreender o conhecimento humano que permeia linguagens naturais e artificiais; e contextualizar essas questões em relação ao idioma português. Dessa forma, o pós-graduado poderá exercer atividades de investigação e desenvolvimento, de prestação de serviços; conferindo qualidade aos processos ligados ao uso de tecnologias da linguagem, que permeiam diversos setores da sociedade.

2.2.4 Destinatários

Os candidatos deverão ser titulares do grau de licenciado, detentores de um currículo escolar, científico ou profissional, cuja capacidade seja reconhecida como apta para a realização do curso de Pós-Graduação em Estudos da Linguagem e Humanidades Digitais. Dessa forma possibilita-se e estimula-se a formação de turmas mistas com base tecnológica e humanas.

Os pós-graduados estarão aptos a desenvolver projetos inter e trans-

disciplinares, centrados em tecnologias da linguagem; desempenhar funções como especialistas em linguagem humana em desenvolvimento de soluções baseadas em tecnologias da linguagem; e atuar na educação, na academia, ou em empresas públicas e privadas.

2.2.5 Metodologia de Ensino e de Avaliação

Nas aulas, que podem decorrer em formato *e-learning* ou presencial, espera-se a participação ativa dos alunos na elaboração de seminários realizados em grupos, constituídos de apresentação oral e entrega dos respetivos relatórios, para cada unidade curricular.

Como bibliografia nuclear, inicial, cobrindo as áreas principais de Humanidades Digitais, suas relações com a língua portuguesa e suas tecnologias, e boas práticas em projetos, sugere-se:

- [McCabe, 2017]
  McCabe, A. (2017). *An Introduction to Linguistics and Language Studies*. Equinox Publishing Limited.

- [Viola, 2023]
  Viola, L. (2023). *The Humanities in the Digital: Beyond Critical Digital Humanities*. Springer Nature.

- [Caseli and Nunes, 2024]
  Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 ed.

– [De Sordi, 2017]
De Sordi, J. O. (2017). *Desenvolvimento de projeto de pesquisa*. Saraiva Educação SA.

– [Gonçalves and Banza, 2013]
Gonçalves, M. F. and Banza, A. P. (2013). *Património Textual e Humanidades Digitais: da antiga à nova Filologia*. CIDEHUS.

– [Castillo Gómez, 2024]
Castillo Gómez, A. (2024). Paleografia, história social de la cultura escrita y revolucion digital. *LaborHistórico*, 10(1).

– [Vieira and Banza, 2022]
Vieira, R. and Banza, A. P. (2022). *Actas da Jornada de Humanidades Digitais do CIDEHUS*. Imprensa da Universidade de Évora.

– [Shneiderman, 2022]
Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

– [Rigolot, 2020]
Rigolot, C. (2020) Transdisciplinarity as a discipline and a way of being: complementarities and creative tensions. *Humanities and Social Sciences Communnications* 7 (100).

Literatura adicional deverá estar alinhada com cada unidade curricular, pensada de forma transdisciplinar entre o conjunto de professores, e

alinhada com os projetos escolhidos, seja por produção científica do próprio projeto identificado, seja por relevância temática em relação à área do projeto.

Para a avaliação, deverão ser considerados os itens assiduidade, participação nas aulas, execução das atividades propostas, bem como o desenvolvimento e a apresentação de seminários e projetos com relatórios escritos. Trabalhos em grupo serão recomendados, estimulando inter e transdisciplinaridade prática e crítica.

2.2.6 Unidades Curriculares

De acordo com os eixos apresentados em 2.1, listamos possíveis unidades curriculares.

- 1º Semestre:
  - Estudos da Linguagem;
  - Introdução a Humanidades Digitais;
  - Multi, Inter e Transdisciplinaridade;
  - Paleografia, Filologia e Digitalização textual.
- 2º Semestre:
  - Evolução da Comunicação Homem-máquina;
  - Introdução a IA, PLN e Modelos de Linguagem;

– Dados abertos, Dados FAIR e Ontologias;

– Projetos em Humanidades Digitais.

Exemplos de questões a abordar na UC de projetos

Os projetos da UC Projetos em Humanidades Digitais podem incorporar questões como as exemplificadas abaixo (lista ilustrativa, não exaustiva):

1) Anotação textual manual e/ou automática;
2) Análise de conteúdo de corpus;
3) Análise de sentimentos em corpus;
4) Preparação de manuscritos até sua digitalização;
5) Ligação de textos com bases de dados como Wikipédia;
6) Proposta de uma ontologia para descrição de meta-dados;
7) *Fairificação*[1] de dados de alguma base ou corpus existente .

2.2.7 Dissertações e Teses

As UCs serão acrescidas de dissertação ou tese para possível obtenção de grau, em nível de mestrado ou doutorado, nesse caso projetos de teor correspondente deverão ser desenvolvidos. Para a realização dessas teses considera-se a relevância de orientação interdisciplinar.

[1]Tornar adequados aos princípios FAIR.

Parte 3

Considerações finais

A linguagem humana atravessa uma fase de transformação profunda de influência tecnológica. O processamento computacional da língua se faz presente no tratamento de problemas em diferentes áreas, incluindo as Humanidades. A investigação em Humanidades Digitais baseadas em fontes textuais faz uso dessas tecnologias em diferentes fases do processo de investigação. Nas fases iniciais enfrentam: a transcrição de manuscritos, o tratamento de textos disponibilizados como imagem, a normalização textual, e a adição de meta-dados, etc. Para as fases posteriores de processamento, a partir de um texto digital preparado, as tecnologias de linguagem podem prestar auxílio a tarefas diversas como: tradução, recuperação e extração de informações, criação de bases de conhecimento, e sua associação a ontologias ou outros dados. No que diz respeito à partilha de material para investigação, é ideal dar atenção especial aos padrões de disponibilização de dados, como os propostos pelos princípios *FAIR* de Dados Abertos, associados a uma descrição semântica bem fundamentada.

As áreas de intersecção entre humanidades e tecnologias requerem um trabalho interdisciplinar. Questões relevantes em diferentes domínios devem ser combinadas. Unem-se, por um lado, os objetivos dos investigadores das áreas de humanidades (linguística e literatura, ciências sociais, história) por outro, os objetivos dos investigadores das áreas tecnológicas que trabalham com o processamento de língua. É preciso trabalhar em conjunto para encontrar, aplicar ou mesmo desenvolver soluções apropriadas para cada tipo de problema, maximizando a correção e o desempenho das soluções encontradas, de acordo com a disponibilidade de recursos.

É crucial que uma perspetiva de IA centrada no ser humano seja levada em consideração [Shneiderman, 2022]. É necessário fornecer interfaces de usuário, adequadas para preparar e acessar as fontes textuais, trabalhar com os dados extraídos e com as ferramentas de análise. Os diversos projetos em curso no CIDEHUS, que consideram coleções distintas, com diferentes objetivos, podem se beneficiar com a troca de experiências e com o uso das mesmas ferramentas para lidar com os textos e seus conhecimentos codificados.

Os métodos usuais de tratamento textual, desenvolvidos em pesquisas de PLN e IA, podem requerer adaptação a diferentes necessidades de investigação, estilo textual, objetivos e também ao uso pretendido e seus usuários. Para que os métodos desenvolvidos sejam de fácil manipulação, o desenvolvimento de novas interfaces se faz necessário.

Reconhecer o desenvolvimento tecnológico da língua é essencial para a o desenvolvimento do seus falantes. A atividade humana é baseada

em trocas que só são possíveis através da comunicação, que hoje envolvem linguagens naturais e artificiais. A tecnologia está presente na busca de informação, nos tradutores, em sistemas de pergunta e resposta, nos robôs conversacionais (*chatbots*), e em diversos aparelhos que funcionam por comandos por voz. Nos dias de hoje, as tecnologias de linguagem estão fortemente presentes nos diversos afazeres quotidianos, com impactos na própria transformação e evolução das línguas.

Nesse documento foram apresentadas as ideias centrais de um plano de investigação que lida com as tecnologias da linguagem em prol das Humanidades Digitais. Essa investigação já encontra-se em curso, em colaboração com uma equipa interdisciplinar do CIDEHUS que reúne linguística, história, informática, turismo e arqueologia e pode atrair ainda mais outras áreas. Apresentou-se conjuntamente uma proposta inicial para um futuro curso interdisciplinar de pós graduação em estudos da linguagem e humanidades digitais.

Referências

[Albuquerque et al., 2024] Albuquerque, G. C., Souza, M., Vieira, R., and Ribeiro, A. S. (2024). Applying event classification to reveal the estado da índia. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 247–254, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

[Banza, 2022] Banza, A. P. (2022). A edição digital da história do futuro, de antónio vieira: arquivo e ferramentas. In *Actas da Jornada de Humanidades Digitais do CIDEHUS*.

[Cameron et al., 2023] Cameron, H. F., Olival, F., and Vieira, R. (2023). Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *Revista LaborHistórico*, 9(1).

[Caseli and Nunes, 2024] Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[Castillo Gómez, 2024] Castillo Gómez, A. (2024). Paleografía, historia social de la cultura escrita y revolución digital. *LaborHistórico*, 10(1).

[Claro et al., 2024] Claro, D. B., Santos, J., Souza, M., Vieira, R., and Pinheiro, V. (2024). Extração de informação. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[Cortes et al., 2024] Cortes, E. G., Vieira, R., and Barone, D. A. C. (2024). Perguntas e respostas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[De Sordi, 2017] De Sordi, J. (2017). *Desenvolvimento de projeto de pesquisa*. Saraiva Educação SA.

[Ferreira et al., 2024] Ferreira, A. P., Garcia, L. D., Dores, M., and Sequeira, O. (2024). Palaeography and diplomatics on the digital humanities route: pathways and proposals. *LaborHistórico*, 10(1).

[Fonseca et al., 2024] Fonseca, E., Vanin, A. A., and Vieira, R. (2024). Resolução de correferência. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[Freitas, 2024] Freitas, C. (2024). Dataset e corpus. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[Gonçalves and Banza, 2013] Gonçalves, M. F. and Banza, A. P. (2013). *Património Textual e Humanidades Digitais: da antiga à nova Filologia*. CIDEHUS.

[Guarino, 1998] Guarino, N. (1998). *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, volume 46. IOS press.

[Guizzardi, 2020] Guizzardi, G. (2020). Ontology, Ontologies and the "I" of FAIR. *Data Int.*, 2(1-2):181–191.

[Lopes et al., 2024] Lopes, R., Magalhaes, J., and Semedo, D. (2024). GlórIA: A generative and open large language model for Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

[Marino et al., 2024] Marino, E., Vieira, R., Baleato, Decoding Sentiments about Migration in Portuguese Political Manifestos (2011, . ., Ribeir, A. S., and Laken, K. (2024). Decoding sentiments about migration in portuguese political manifestos (2011, 2015, 2019). In *Digital Humanities and Natural Language Processing Workshop,DHandNLP@PROPOR, no prelo*.

[McCabe, 2017] McCabe, A. (2017). *An Introduction to Linguistics and Language Studies*. Equinox Publishing Limited.

[Mittal and Garg, 2020] Mittal, R. and Garg, A. (2020). Text extraction using ocr: a systematic review. In *2020 second international conference on inventive research in computing applications (ICIRCA)*, pages 357–362. IEEE.

[Nair and Jeeven, 2004] Nair, S. S. and Jeeven, V. (2004). A brief overview of metadata formats. *DESIDOC Journal of Library & Information Technology*, 24(4).

[Paes et al., 2024] Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos de linguagem. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN.

[Rigolot, 2020] Rigolot, C. (2020). Transdisciplinarity as a discipline and a way of being: complementarities and creative tensions. *Humanities and Social Sciences Communnications*, 7(100).

[Rodrigues et al., 2023] Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt. In *EPIA Conference on Artificial Intelligence*. Springer.

[Sacramento and Souza, 2021] Sacramento, A. d. S. B. and Souza, M. (2021). Joint event extraction with contextualized word embeddings for the portuguese language. In *Brazilian Conference on Intelligent Systems*. Springer.

[Santos and Vieira, 2021] Santos, I. and Vieira, R. (2021). Semantic information extraction in archaeology: Challenges in the construction of a portuguese corpus of megalithism. In *15th International Conference on Metadata and Semantics Research, Springer Communications in Computer and Information Science Series, Vol. 1537.*

[Santos et al., 2024] Santos, J., Cameron, H. F., Olival, F., Farrica, F., and Vieira, R. (2024). Named entity recognition specialised for Portuguese 18th-century history research. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

[Santos et al., 2019] Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., and Vieira, R. (2019). Assessing the impact of contextual embeddings for portuguese named entity recognition. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE.

[Shneiderman, 2022] Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

[Souza et al., 2024] Souza, F., Nogueira, R., and Lotufo, R. (2024). Bert models for brazilian portuguese: Pretraining, evaluation and tokenization analysis. *Appl. Soft Comput.*, 149(PA).

[Van Strien et al., 2020] Van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the

impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020) - Volume 1*. SCITEPRESS Publications.

[Vieira and Banza, 2022] Vieira, R. and Banza, A. P. (2022). *Actas da Jornada de Humanidades Digitais do CIDEHUS*. Imprensa da Universidade de Évora.

[Vieira et al., 2021] Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., and Santos, I. (2021). Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

[Viola, 2023] Viola, L. (2023). *The Humanities in the Digital: Beyond Critical Digital Humanities*. Springer Nature.

[Wilkinson et al., 2016] Wilkinson, M., Dumontier, M., and Aalbersberg, e. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.