Capítulo 27

PLN e Humanidades Digitais

Renata Vieira
Helena Freire Cameron
Fernanda Olival
Fatima Farrica
Maria José Bocorny Finatto
Ana Paula Banza
Ana Sofia Ribeiro
Cassia Trojahn

Publicado em: 13/03/2024

27.1 Introdução

A área de humanidades digitais (HD) tem ganhado força e adeptos nas últimas décadas, em paralelo com o desenvolvimento de ferramentas digitais que ampliam as possibilidades de armazenamento, acesso e processamento de dados. Essas capacidades estendem os horizontes de atuação de pesquisadores, permitindo a captura, organização e análise de um volume muito grande de dados. Com esta interação, que pode ser classificada de tecnológica, as humanidades ganham visibilidade, atravessam fronteiras disciplinares e enfrentam desafios sem precedentes.

Este capítulo apresenta as bases comuns para projetos na área de HD, relacionados ao Processamento de Linguagem Natural (PLN), em particular sobre fontes textuais com grafia pré-contemporâneas, o que imprime uma maior complexidade no processamento. Além disso, nosso enfoque é a relação entre as HD e o PLN no contexto da língua portuguesa, entendendo que o alcance da área de HD vai muito além da língua, e do escopo das questões e projetos aqui mencionados.

Na área de HD, relativamente aos trabalhos baseados em fontes textuais, encontramos uma grande variação, tanto nos períodos históricos das fontes, no seu suporte (manuscritos em papel, impressos, fotografados, etc), como no seu estágio de digitalização, que pode variar entre imagens digitais, textos em PDF e textos digitalizados em outros formatos. Todas essas variações adicionam esforços extras de processamento. Desta forma, apresentaremos um panorama geral e discutiremos os requisitos de preparação e organização das fontes, objetos de análise que depois de transcritas e digitalizadas, podem ser submetidas a processamentos mais avançados. O objetivo é mostrar não apenas como o PLN é útil e relevante nesse domínio, mas também como a área de HD é rica em despertar novas questões para o desenvolvimento do PLN.

Este capítulo está organizado da seguinte maneira. Na Seção 27.2 apresentamos os tópicos e os desafios relacionados à preparação das fontes, desde a digitalização até a anotação. Na Seção 27.3, discutimos os processos de transformação de fontes em dados, nos quais várias tarefas conhecidas de PLN (discutidas em outros capítulos deste livro) podem ser empregadas. Esses processos de transformação de textos em dados organizados possibilitam ao pesquisador realizar análises diferenciadas sobre textos, por exemplo, ao agrupar tipos específicos de informação, ou rapidamente quantificar fenômenos observáveis. Na Seção 27.4, apresentamos alguns projetos relacionados às HD, em especial aqueles desenvolvidos pelas autoras. São projetos que apresentam diferentes fontes e objetivos de pesquisa, mas nos quais o PLN se faz presente em vários níveis, desde a preparação e a anotação até a extração, organização e partilha da informação. Por fim, apresentamos nossas considerações finais na Seção 27.5.

27.2 Preparação das fontes

Trabalhos em HD que se baseiam em textos podem se beneficiar das atuais técnicas de Inteligência Artificial (IA) e PLN para um melhor acesso às fontes. Os textos em HD podem ter relevância por suas características históricas e literárias, mas também podem estar relacionados a estudos sociais ou de outra natureza. Na verdade, é difícil distinguir os limites das HD, mas se pode dizer que a área está ligada ao emprego e à compreensão de novas maneiras de se desenvolver pesquisa, com base em recursos digitais e computacionais. Em relação ao processamento da língua, essa influência tecnológica se observa em uma variedade de processos referentes ao tratamento das fontes e sua informação, tais como:

- **Digitalização de documentos**: preparação no nível mais básico, para o tratamento computacional, e assim possibilitar a sua leitura por meio de programação.
- **Adição de metadados**: inserção de níveis de informação extra-textuais, que podem ser importantes para a estruturação e o armazenamento, e também para o acréscimo de interpretações sobre o conteúdo de uma coleção de interesse. Os metadados podem trazer informações de contexto, como autores, datas, locais de produção, origem, ou ainda informações como volume, páginas, etc. Ou podem ser de análise linguística: morfológica, sintática ou semântica (por exemplo, identificando entidades ou eventos).
- **Normalização**: a normalização para uma língua padrão e contemporânea pode ser necessária em situações como fontes históricas, textos de redes sociais que usam muitas abreviações, símbolos associados à emoções, ou comentários como *hashtags*, etc.
- **Anotação de *corpora***: muitas vezes o estudo de uma fonte ou *corpus* necessita a adição de informações extras sobre o registro, que são representadas como metadados (mencionado acima). 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), quer para a avaliação da correção da anotação realizada por máquina.

Os métodos usuais de tratamento textual, desenvolvidos em pesquisas de PLN e IA, podem requerer adaptação a diferentes necessidades de investigação, estilo textual, objetivos

da pesquisa e também ao uso pretendido e seus usuários. Idealmente, são necessárias novas interfaces para que os métodos desenvolvidos sejam usados de forma facilitada e fora do contexto das estruturas de programação. Apresentaremos, a seguir, esses passos iniciais para projetos de HD que lidam com fontes textuais e, posteriormente, discutiremos sobre as técnicas de PLN.

27.2.1 Digitalização

Muitos projetos na área de HD necessitam lidar com a digitalização de manuscritos originais. A área de paleografia, em particular, lida com a leitura e transcrição de versões manuscritas para um suporte atual. Mesmo registros em suportes muito antigos, como a escrita cuneiforme, podem ser transcodificados para o digital (Liu; Hearne; Conrad, 2016), e a partir daí, podem passar por processos computacionais.

A paleografia, deste modo, também se desenvolveu de habilidade tradicional do intelecto, e manual, para uma prática digital (Ackel, 2021). Algumas ferramentas de transcrição, como por exemplo o Transkribus[1] ajudam na digitalização de manuscritos (Kahle et al., 2017), através de um sistema de algoritmos HTR (*Handwritten Text Recognition*).

Alguns documentos impressos podem necessitar a aplicação de tecnologia OCR (*Optical Character Recognition*). A qualidade da saída de sistemas OCR varia muito dependendo do formato e da qualidade da entrada. É um problema básico, mas não totalmente resolvido (Strien et al., 2020). Por exemplo, nas primeiras levas de digitalização de acervos arquivísticos e bibliotecários, muitas digitalizações foram feitas em formatos de imagem, sem a preocupação de possibilitar a pesquisa dos seus conteúdos de texto. Por meio de diferentes ferramentas de transcrição automática com OCR, como o eScriptorum[2], podemos fazer essa transcrição de imagem para texto de forma automática.

Muitas vezes, as saídas de OCR precisam passar por processamento extra de correção, por vezes manual, para tornar a fonte adequada para as próximas etapas. Esse processamento se faz necessário para que outras ferramentas, como as de anotação de texto, tenham maior eficácia na aprendizagem dos *tokens* anotados. É preciso preparar os textos, por exemplo, retirando números de páginas, textos de cabeçalhos ou rodapés, e evitar a translineação/hifenização de palavras. Se o texto a ser anotado for composto por um conjunto de textos, pode ser necessário que estejam devidamente separados. Outro problema que pode emergir são os textos organizados em mais de uma coluna nas páginas, pois será preciso ordená-los numa segmentação textual contínua. Essas etapas devem ser consideradas em projetos de HD envolvendo textos e PLN e é, portanto, necessário antever-se o tempo e recursos necessários para o seu desenvolvimento.

27.2.2 Metadados

Uma vez digitalizada a fonte ou *corpus* de estudo, é necessário pensar na organização dos seus metadados. O material digital deve ser bem descrito, ou seja, deve conter a informação sobre a qual acervo pertence, identificar unicamente cada arquivo e, quando pertinente, associar autoria, data e outros elementos pertinentes. Os metadados também podem descrever a estrutura do documento, quando necessário, identificar volumes, capítulos,

[1] https://www.transkribus.org
[2] https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/

páginas, por exemplo. É importante, ainda, separar os metadados dos dados originais. Para materiais transcritos, por exemplo, devem ser identificados os cabeçalhos, notas de rodapé, numeração de páginas ou comentários adicionados aos originais. Esta organização e distinção dos elementos extra-textuais são essenciais para garantir o bom processamento posteriormente. Os metadados são essenciais para conectar uma fonte a outras e as conectar aos dados abertos ligados (Nair; Jeeven, 2004).

Há vários padrões de metadados bastante difundidos. Por exemplo, a Text Encoding Initiative (TEI) é um consórcio que desenvolve e mantém coletivamente um padrão para a representação de textos em formato digital. O seu principal resultado é um conjunto de diretrizes que especifica métodos de codificação para textos legíveis por máquina, principalmente em ciências humanas, ciências sociais e linguística[3]. Outro exemplo é o grupo de trabalho sobre "Publicação Sustentável de Metadados", ativo na rede europeia da *Digital Research Infrastructure for the Arts and Humanities* (DARIAH) desde 2016. Esse grupo tem como objetivo apresentar informações sobre a reprodução e o compartilhamento de metadados entre institutos e pesquisadores no campo das ciências humanas e sociais.

Vários trabalhos trataram da representação de metadados por meio de vocabulários semânticos, para descrever quer recursos digitais, como vídeos, imagens, páginas web, quer recursos físicos, como livros ou obras de arte. Esse vocabulários podem ser usados para fornecer interoperabilidade para o uso de metadados na nuvem de dados ligados. Exemplos são o Dublin Core[4], VoID (Vocabulary of Interlinked Datasets[5]), schema.org, DCAT (Data Catalog Vocabulary[6]) e PROV-O (Provenance Ontology[7]). Em particular, o DCAT é um vocabulário projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web. Permite a descrição de conjuntos de dados e serviços, usando um modelo e um vocabulário padrão que facilitam o consumo e a agregação de metadados de vários catálogos. Os metadados agregados do DCAT podem servir como um arquivo de manifesto[8] como parte do processo de preservação digital.

27.2.3 Normalização

Em HD, as fontes textuais, quer manuscritas, quer impressas, sendo de uma época anterior ao estágio atual da língua, apresentam variações ortográficas ou morfossintáticas, não só em relação ao padrão atual como também dentro da mesma época.

A variação gráfica no século XVIII, por exemplo, é muito expressiva, quer em textos manuscritos, quer em impressos, especialmente ao nível da representação das vogais e de ditongos nasais, por exemplo "am-ão", em "fizerão", ou "oens-ões", em "embarcaçoens", etc. Também as consoantes duplas como "ll-l", em "delles", "tt-t", em "Cratto", entre outras, têm uma elevada frequência nesses textos. Note-se, também, a existência do fenômeno das chamadas grafias cultas ou pseudo-etimológicas, que já não integram o cânone atual da língua portuguesa. Estas podem trazer uma maior complexidade ao processamento lexical, como "th-t", em "athé", "ch-c" em "Christo" e "y-i", por exemplo na palavra "Rey". A

[3] https://tei-c.org
[4] https://www.iso.org/standard/71339.html
[5] http://vocab.deri.ie/
[6] https://www.w3.org/TR/vocab-dcat-3/
[7] https://www.w3.org/TR/prov-o/
[8] *Manifest file*, arquivo que contém metadados para um grupo de arquivos que são parte de um conjunto ou de uma unidade coerente.

palavra "coro", que frequentemente é registada como "choro" em textos do século XVIII, é um exemplo da complexidade da variação gráfica deste período.

Do ponto de vista do leitor, a abordagem destes documentos requer um forte conhecimento do contexto para a sua desambiguação e exige, frequentemente, conhecimentos linguísticos e históricos para que possa desvendar o conteúdo de forma adequada. A variação gráfica tem uma elevada incidência nos documentos desta época. Grande parte das palavras tem pelo menos uma variante gráfica e algumas palavras têm múltiplas num mesmo texto. A título ilustrativo, veja-se a palavra "circunstância", com nove variantes gráficas recolhidas num mesmo *corpus* textual (Cameron; Olival; Vieira, 2023a).

Do ponto de vista do PLN, a variação, quer pela sua grande incidência, quer pela sua imprevisibilidade, vem trazer uma perturbação acrescida. Variantes de uma mesma palavra são tratadas como se fossem unidades lexicais diferentes, diminuindo a eficácia do processamento. Para que se possa processar o texto de forma mais célere e eficaz, frequentemente têm de ser desenvolvidas tarefas intermediárias de normalização, quase sempre de forma manual. Unificam-se as variantes gráficas em torno de um lema, normalmente a forma existente na ortografia padrão atual. Essa tarefa, morosa e minuciosa, muito ganharia com processos automáticos ou semi-automáticos de normalização sistemática de textos desses estágios da língua portuguesa. Essa tarefa não só os tornaria aptos para futuros processamentos, mas também permitiria que leitores não historiadores ou não linguistas pudessem ter acesso mais facilmente a textos valiosos culturalmente e de valor patrimonial, com interesse para várias áreas do saber. Compreender essas diferenças e conseguir traduzir ou associar escritos antigos aos padrões atuais é um passo importante para outros níveis de processamento (Cameron; Gonçalves; Quaresma, 2020).

Outra forma de mitigar as variantes de escrita é criar modelos de linguagem que incluam *corpora* de outros períodos de tempo, com suas variantes de ocorrência natural, numa fase adicional de treinamento (ajuste por *fine tuning*), para adaptar o modelo às variantes (Arevalo; Fonteyn, 2021). Em *corpora* históricos transcritos, a presença de vários critérios de transcrição pode dificultar a operação. Além disso, há a dificuldade em reunir um volume substancial de textos de uma mesma época que permita aprendizagens eficientes e com elevada acuidade.

27.2.4 Anotação de textos

A anotação de textos é um processo que por vezes pode ser complexo, necessitando de equipes interdisciplinares, de modo a prever todas as vertentes. A anotação requer a identificação de categorias de interesse para o estudo realizado, isso implica em impor alguma objetividade sobre a língua, sendo que a interpretação é fundamentalmente um processo subjetivo.

A anotação textual ocorre em diferentes níveis, sintático, semântico lexical, e discursivo. Como para qualquer outra área, também nas HD, é útil identificar elementos de interesse para embasar diferentes estudos. Geralmente a anotação sintática permite um tipo de análise diferenciada da anotação semântica. Estudos linguísticos de filologia podem se beneficiar das anotações de nível morfológico (lexical estrutural) e sintático (sentencial), como poderemos conferir em exemplos de projetos apresentados nas Seções 27.4.1 e 27.4.4.

A anotação de entidades é frequentemente adotada, uma vez que se trata de uma tarefa com desenvolvimento bastante avançado, com utilidade para estudos de literatura, história,

geografia, entre muitos outros campos do conhecimento. Conforme a complexidade do fenômeno e a respectiva disponibilidade de recursos, a anotação pode ser manual ou automática.

Identificar personagens, instituições ou os locais mencionados em documentos históricos ou obras literárias possibilita uma série de análises. Quase sempre permite uma resposta mais consistente a perguntas básicas, mas essenciais em qualquer estudo rigoroso, como são as associadas a questões como "quem?","o quê?", "onde?". Isso poderá ser verificado nas Seções 27.4.2, 27.4.3 e 27.4.5. No processo de anotação, é fundamental respeitar o contexto nas decisões que o anotador deve tomar, pois há muitos casos que se prestam a interpretação ambígua.

27.3 Transformação de textos em dados

Naturalmente, com a evolução dos estudos em HD, são produzidos dados diferenciados, mais elaborados, mais numerosos, e úteis aos investigadores em humanidades. Amplia-se, assim, a capacidade de análise, pela possibilidade de trabalhar com um volume maior de informação e pela capacidade de organizar essa informação de forma mais ágil e rápida em estruturas bem definidas. O PLN atua como 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 re-uso 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 pesquisadores das respectivas áreas. Isso é necessário para o desenvolvimento das adaptações de ferramentas a diferentes objetivos e para a construção de interfaces adequadas ao seu uso. É, de fato, uma elaboração interdisciplinar. É importante aliar os interesses dos pesquisadores em humanidades às possibilidades mais atuais e eficientes de obtenção de informação por meio da aplicação de tecnologias da linguagem, especialidade dos pesquisadores em PLN. Discutiremos a seguir os elementos, mencionados acima, concernentes à transformação de textos em dados.

27.3.1 Extração de informação

Entre as subáreas de PLN, 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). O capítulo 17 deste livro apresenta em detalhes essa área do PLN. Enquadram-se nela as tarefas de reconhecimento de entidades nomeadas e de identificação e classificação de eventos, aplicadas, por exemplo, nos projetos descritos nas Seções 27.4.2, 27.4.3 e 27.4.5 para o estudo de fontes históricas.

O reconhecimento de entidades nomeadas possibilita a identificação de personagens importantes 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 pesquisadores. A identificação de eventos e sua respectiva classificação podem ajudar a abordar zonas e contextos significativos, como, por exemplo, no projeto *Monsoon* (Seção 27.4.3) que realiza uma análise dos eventos inerentes a conflitos na coleção "Livro das Monções".

Em 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, resolução de correferências, e sistemas de respostas a perguntas.

Uma outra maneira de fazer uso de ferramentas prontas, em textos não normalizados, com características diferenciadas do padrão contemporâneo, seria adaptar os modelos de linguagem já treinados. Essas adaptações podem ser feitas para datas históricas ou para a diversidade de mídias. Uma adaptação pode melhorar a performance dos sistemas, uma vez que os grandes modelos são treinados em versões contemporâneas da língua. Não conhecemos alguma adaptação de modelos para o português histórico, mas em breve deverão estar disponíveis, pois esta é uma área em rápida evolução.

27.3.2 Representação de conhecimento e ontologias

Ontologias são artefatos conceituais interpretados com diferentes nuances em filosofia, lógica e sistemas de informação. A ideia central é representar conceitualizações de forma estruturada. Em HD, a nuance mais comumente relacionada é aquela empregada em sistemas de informação, ontologias como estruturas conceituais que podem ajudar na organização dos dados e na comunicação entre sistemas. O nível de detalhe de uma ontologia pode variar entre estruturas taxonômicas, definição de instâncias, relações e axiomas lógicos. Elas auxiliam a esclarecer (e combinar) diversas conceitualizações que compõem diferentes disciplinas.

Ontologias têm sido usadas em diversos domínios do conhecimento em diferentes tarefas:

- organizar e anotar grandes quantidades de dados (anotação semântica);
- integrar dados de diversas fontes (integração semântica);
- representar conhecimento de domínios complexos;
- dar ancoragem para o raciocínio automático:
- suportar buscas em grandes quantidades de dados (busca semântica).

Em HD, além das tarefas citadas acima, ontologias têm sido utilizadas para representação de metadados, pois, geralmente, é necessário estender ontologias de domínio existentes ou desenvolver novas para atender às especificidades de cada *corpus* ou fonte. Essas descrições formais permitem a integração de dados de múltiplas fontes, de maneira independente de software e de esquema. Um passo essencial para melhorar a qualidade dos dados é

usar vocabulários padronizados e ontologias para representação de dados e metadados (Guizzardi, 2020).

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

27.3.3 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 e *Reuse*/Reutilização) (Wilkinson; Dumontier; Aalbersberg, 2016). Esses princípios correspondem a um conjunto de 15 recomendações que visam facilitar a reutilização de dados por humanos e máquinas. Eles são independentes de domínio e podem ser implementados principalmente através de:

- (F) atribuição de identificadores exclusivos e persistentes a conjuntos de dados, descrevendo-os com metadados ricos que permitem sua indexação e descoberta;
- (A) uso de protocolos abertos e de padrões para acesso a conjuntos de dados;
- (I) uso de linguagens formais e de vocabulários padronizados *FAIR* para representar (meta)dados;
- (R) uso de metadados ricos sobre licença de uso, proveniência e qualidade de dados.

Portanto, o primeiro passo para o cumprimento dos princípios *FAIR* é atribuir metadados aos conjuntos de dados e definir esquemas precisos de metadados. De fato, 12 dos 15 princípios *FAIR* (Wilkinson; Dumontier; Aalbersberg, 2016) se referem a metadados. Para dar um passo adiante no aprimoramento *FAIR* dos dados, os esquemas de metadados devem se basear em modelos semânticos (ou seja, ontologias) para uma representação de metadados mais rica (Guizzardi, 2020). Graças à sua capacidade de tornar os tipos de dados explícitos, em um formato que pode ser processado por máquinas, as ontologias são essenciais para tornar os dados *FAIR*, mesmo para dados já publicados na Web (Jacobsen et al., 2020).

A ideia dos dados ligados é que os esforços possam ser reaproveitados e as bases enriquecidas com dados já identificados e organizados por outras iniciativas. A *open linked data cloud* (LOD)[9] fornece uma visão geral dos conjuntos de dados vinculados que estão disponíveis na Web. Representa um grafo de conhecimento, composto por diversos padrões abertos, como URI, URL, HTTP, HTML, RDF, a linguagem de consulta SPARQL, entre outros. Esse padrões são mantidos por curadores de dados amadores e profissionais na indústria e na academia[10].

[9] https://lod-cloud.net
[10] https://www.w3.org/wiki/LinkedData

Por exemplo, a *Europeana* é uma biblioteca virtual desenvolvida por países da União Europeia (Borin; Donato, 2023; Coneglian; Santarem Segundo, 2017). O protótipo agrega milhões de itens digitais, todos eles em domínio público, e está ligado ao LOD Cloud[11]. Em (Koho et al., 2021) é proposta uma infraestrutura semântica compartilhada, baseada na ideia de representar conceitos de guerra como uma sequência espaço-temporal de eventos, nos quais participam soldados, unidades militares e outros atores. O esquema de metadados usado é uma extensão do CIDOC-CRM, complementado por várias ontologias de domínio histórico militar, e é ligado ao LOD Cloud.

Em termos de avaliação, várias estruturas foram propostas para avaliar o grau de *FAIRness* de um determinado objeto digital (Sun; Emonet; Dumontier, 2022). Em vários deles, a avaliação é realizada respondendo a um conjunto de perguntas, também chamadas de métricas ou indicadores, ou preenchendo uma lista de verificação[12], como o *FAIRshake* (Clarke et al., 2019) ou o *FAIR Data Maturity Model* (FAIR Data Maturity Model Working Group RDA, 2020).

Outros autores (Devaraju et al., 2020; Wilkinson; Dumontier; Sansone, 2019) propuseram abordagens automatizadas. Recentemente, além de avaliar o grau de *FAIRness* dos dados, as propostas abordaram a avaliação de vocabulários e ontologias (Cox, 2021; Garijo; Poveda-Villalón, 2020).

Nesta seção discutimos alguns elementos do caminho de construção de bases de pesquisa a partir de textos. Muitos projetos na área de HD são desenvolvidos com o objetivo de mapear informações textuais volumosas, dispersas, ou ainda não totalmente exploradas em novas descobertas. Uma vez mapeadas e organizadas, elas podem estar disponíveis para outros investigadores. As áreas de extração de informação, construção de ontologias e dados ligados caminham em paralelo, embora esforços interdisciplinares ainda sejam necessários para sua integração. Na próxima seção serão apresentados exemplos de projetos que caminham nesse sentido.

27.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa

Apresentamos aqui exemplos de trabalhos envolvendo HD e PLN com foco em língua portuguesa. Esta não é uma apresentação completa; há diversos outros projetos e trabalhos importantes e reconhecidos nesse domínio. Em especial, damos atenção aqui aos projetos relacionados com as autoras deste capítulo, e que colaboram no contexto do Laboratório Chronos[13], o Laboratório de Humanidades Digitais do Centro Interdisciplinar de História Culturas e Sociedades, CIDEHUS, da Universidade de Évora, Portugal. São, em geral, projetos atuais em desenvolvimento, com exceção do primeiro, que tem uma relevância histórica no desenvolvimento dessa área no Brasil.

27.4.1 O *corpus* Tycho Brahe

O *corpus* histórico português Tycho Brahe (Sousa, 2014) é um *corpus* eletrônico de textos escritos em português por autores nascidos entre 1380 e 1978. O objetivo principal

[11] https://www.europeana.eu/en
[12] https://fairassist.org/
[13] https://sites.google.com/view/hdlabcidehus

do projeto foi possibilitar, de forma ampla, a recuperação de informações filológicas e linguísticas dos textos. Esse pode ser considerado um dos trabalhos pioneiros na área de *corpus* histórico, com adição de anotação linguística e que tenha envolvido recursos computacionais na sua criação, sendo um importante contributo para essa área (Britto; Finger; Galves, 2002; Finger, 2000).

No site do projeto[14], 95 textos se encontram disponíveis para pesquisa. Os textos contam com anotação linguística em dois níveis: anotação morfológica *(part of speech)* (60 textos, total de 2.204.889 palavras); e anotação sintática (31 textos, total de 1.311.834 palavras). Estão disponíveis ainda versões sem anotação. A coleção engloba o português brasileiro e o europeu, contendo, por exemplo, a Gazeta de Lisboa e o Jornal da Bahia.

27.4.2 As Memórias Paroquiais

As Memórias Paroquiais constituem uma das principais fontes sobre o Portugal metropolitano dos meados do século XVIII. São, aliás, das mais citadas pelos historiadores. A coleção é composta pelas respostas dos párocos a um inquérito lançado em 1758, três anos depois do grande sismo de 1755. Incluía 60 perguntas sobre o território da freguesia, pelo que reúne uma plêiade diversificada de informações, desde elementos administrativos e demográficos aos recursos existentes: piscícolas, minerais, aquíferos com propriedades específicas, produções agro-pecuárias, etc. Interessam tanto ao historiador como ao botânico ou ao sismógrafo, para referir apenas alguns dos campos abrangidos pela observação dos párocos informantes.

Estas respostas, encadernadas no século XIX, fez com que os volumes passassem a ser conhecidos como "Dicionário Geográfico de Portugal". Aliás, a iniciativa de 1758 também tinha em vista dar continuidade a um projeto de dicionário com esse perfil, do Padre oratoriano Luís Cardoso (1697-1769), e que o terremoto de 1755 interrompera. A atual designação só se impôs no limiar do século XX (Olival; Cameron; Vieira, 2022).

Os originais manuscritos, num total de 43 volumes e um de índices, encontram-se na Torre do Tombo, desde o século XIX. Entre 1993 e 2003, os textos foram microfilmados e em 2005 esses rolos de 35 mm[15], em preto e branco, foram digitalizados e disponibilizados *online*.

Por volta de 2007-2008, o CIDEHUS iniciou a tarefa de transcrever as Memórias das localidades a Sul do Tejo. O trabalho foi concluído em 2023 em um projeto colaborativo. Por vezes, até aproveitou transcrições já publicadas, sempre que foi possível entrar em contato com o Autor(a) da edição. Parte desse trabalho encontra-se disponível online, em acesso aberto, e com licenças *Creative Commons*, que facilitam a reutilização.

Nos últimos dois anos começaram a ser aplicadas técnicas de PLN ao conjunto transcrito. Principiou-se por anotar categorias simples e básicas: pessoas, locais e organizações (Vieira et al., 2021); já em 2023, usaram-se categorias de anotação mais complexas e com várias subdivisões. Teve-se em vista responder de modo mais adequado às necessidades do historiador e de outros cientistas ou pessoas interessadas. Também foi efetuado um ensaio de reconhecimento de entidades nomeadas, usando sistemas previamente desenvolvidos e que foram calibrados para esse efeito (Santos et al., 2024). Baseavam-se em técnicas de aprendizagem de máquina e modelos de linguagem.

[14]https://www.tycho.iel.unicamp.br/corpus/

[15]Agradece-se à Dra. Anabela Ribeiro, do Arquivo Nacional da Torre do Tombo, estas informações.

Está ainda em curso a normalização da grafia do texto, quer em versão com léxico explicativo para o grande público, quer em versões para processamento, trabalhando-se na possibilidade de normalizar de forma automatizada (Cameron; Olival; Vieira, 2023a).

Nos próximos anos, será dada continuidade à tarefa de transformar estes textos em dados confiáveis e de ligá-los a outros repositórios de conhecimento.

27.4.3 O projeto MONSOON

Extração de eventos é uma tarefa bastante conhecida de PLN, com recursos desenvolvidos para o português (Sacramento; Souza, 2021). O desafio, neste caso, é adaptar essas ferramentas à língua portuguesa do século XVII, esforço que está sendo explorado no projeto *Monsoon*: o Estado da Índia Habsburgo em perspectiva digital (1580-1640).

Este projeto pretende estudar as dinâmicas internas desse espaço macro que constituía o Estado da Índia português, combinando o estudo da presença formal portuguesa com a presença de comunidades portuguesas que se instalaram na Ásia fora do território administrado pela Coroa portuguesa, buscando relacionar estas duas esferas com o Outro – seja o europeu ou o asiático ou africano. Por outro lado, o projeto tem o objetivo de criar uma metodologia de análise semi-automática que possa ser replicada em coleções documentais semelhantes de outras potências europeias (Ribeiro, 2022).

Este projeto tem como base documental a correspondência trocada entre o monarca português e as suas instituições com o vice-rei da Índia. Este conjunto ficou conhecido como Livros das Monções, dado que era na altura da monção no Oceano Índico que os navios portugueses podiam partir de Goa para Portugal ou podiam chegar ao Índico. Só no regime climático de monções existem ventos no Índico que permitem que navios sem motor aí circulem. Este conjunto documental, dividido entre o Arquivo Nacional Torre do Tombo, em Lisboa, Portugal, e o Arquivo Histórico do Estado de Goa, em Pangim, Índia, possui informações sobre os mais variados aspectos da presença portuguesa na Ásia e abarca uma geografia que vai desde Moçambique e toda a costa oriental africana até ao Japão.

Este *corpus* documental é composto por um grande número de volumes, sendo que nenhum deles possui um índice sobre os assuntos tratados. Por meio da extração automática de eventos, os historiadores podem facilmente ter acesso a passagens relevantes a partir de uma palavra relacionada com uma determinada ação. É possível identificar e classificar semanticamente eventos, o que nos permite ter uma percepção clara dos temas abordados nesta correspondência, de forma eficiente e rápida. A partir da aplicação de extração de eventos numa amostra destas cartas, compreendemos que, em determinada cronologia, havia no Oceano Índico uma conjuntura generalizada de conflito, quer devido à concorrência de outras potências ultramarinas europeias, como a Holanda e a Inglaterra, quer entre os portugueses e outras entidades políticas asiáticas (Albuquerque et al., 2024).

Através do sistema de extração de eventos, temos acesso a estatísticas de classificação que nos permitem uma percepção mais evidente dos assuntos representados em cada carta e assim analisar conjunturas históricas. Isso, no que diz respeito à metodologia da análise histórica, é algo relativamente novo.

Por outro lado, este projeto utiliza também a anotação de entidades nomeadas de quatro tipos: pessoas, locais, grupos de pessoas e instituições. Nos próximos passos deste projeto pretendemos conjugar as entidades extraídas por meio de anotação de texto e ligá-las aos eventos extraídos no texto, podendo assim articular agentes, instituições e locais às ações

em que estão envolvidos.

27.4.4 A História do Futuro

Nos projetos em curso sobre a "História do Futuro" (Banza, 2022), de Padre António Vieira, após alguns testes com a ferramenta de transcrição de manuscritos Transkribus (Kahle et al., 2017), concluiu-se que, tratando-se de um texto relativamente curto e dadas as características do manuscrito, de leitura muito difícil, em parte devido às más condições do suporte, o treino necessário para uma utilização rentável da ferramenta não se justificaria neste caso, ao contrário de projetos envolvendo um grande volume de textos, uma vez que seria sempre necessária uma correção manual. Assim, o texto foi lido e transcrito manualmente para suporte digital, aplicando-se apenas posteriormente determinadas ferramentas de análise, de acordo com os objetivos pretendidos.

A primeira transcrição, feita a partir da leitura paleográfica do manuscrito, sendo maximamente conservadora, mantém um número muito significativo de traços do português seiscentista, e do português do Padre António Vieira em particular, que dificultam muito ou impossibilitam mesmo a aplicação da maioria das ferramentas de PLN. Assim, optou-se por, a partir da primeira, realizar uma segunda transcrição, igualmente em formato , mas normalizadora. Dadas as características do texto, a normalização foi, mais uma vez, feita manualmente, depois de definidos critérios rigorosos que, sem desvirtuarem as principais características do texto e da sua época, o tornassem apto a processamento computacional e a aplicação de ferramentas automáticas.

Sobre a versão normalizadora, tem sido possível aplicar com êxito algumas ferramentas e testar outras. Inicialmente, constituiu-se um *corpus* não lematizado, que foi anotado lexicalmente (permitindo, por exemplo a elaboração de um léxico da obra) e morfologicamente (o que permitirá rastrear automaticamente alguns aspetos inovadores na obra de Vieira neste domínio, como é o caso do uso dos pronomes clíticos), com recurso à ferramenta LX-Tagger[16], disponível no repositório PORTULAN-CLARIN[17].

Nesses domínios, a anotação automática, verificada depois manualmente, revelou-se particularmente eficaz. Elaboraram-se, ainda, com auxílio da ferramenta AntConc[18], listas lexicais, alfabéticas, por frequência descendente, ou ordenadas pelo final de palavra, permitindo ver a disponibilidade do sistema sufixal. O *corpus* contém ainda um valioso intertexto latino, que foi processado manualmente de modo a constituir um sub-*corpus* autônomo.

Ainda no domínio da análise linguística, prevê-se a utilização do esquema de Dependências Universais, por exemplo, para a análise morfossintática, cujos testes têm revelado um desempenho bastante razoável, apesar de, como nos demais casos, não se poder excluir a revisão manual. Encontram-se também ainda em fase de avaliação ferramentas que permitam identificar similaridades semânticas (Lopez-Gazpio et al., 2017) entre o texto consensualmente aceito como "História do Futuro" e outros, fisicamente distintos, mas conceitualmente idênticos. Neste aspecto, as ferramentas presentemente disponíveis têm-se revelado pouco adequadas, quer ao tipo de texto, quer ao tipo de similaridades que se procura detectar. O objetivo será, neste caso, adaptar os modelos atuais à"História do

[16]http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXTagger.html
[17]https://portulanclarin.net
[18]https://www.laurenceanthony.net/software/antconc/

Futuro", conseguindo, essencialmente, mais eficiência, em termos de precisão e cobertura. Ainda assim, algumas ferramentas já disponíveis, como por exemplo, para o reconhecimento de entidades nomeadas ou para a análise de sentimentos, úteis para a análise histórico-cultural e literária da obra, poderão necessitar de uma menor adaptação.

27.4.5 Manuais médicos do século XVIII

Com a digitalização de acervos bibliográficos de bibliotecas físicas, impulsionada pelas HD (Finatto, 2023), tornou-se mais fácil acessar obras médicas antigas impressas em português. Todavia, tais obras precisam ter seus conteúdos sistematizados para que possam ser compreendidas e situadas em relação ao conhecimento médico e científico atual.

Mesmo com a digitalização e com as transcrições dos textos, conforme sua apresentação original, a interpretação e sistematização dessa informação requer consulta a toda uma série de materiais de apoio: compêndios de História da Ciência e da Medicina, edições filológicas, bases de dados com textos do português de diferentes épocas e vários tipos de dicionários. Assim, no âmbito do projeto "Humanidades digitais e conhecimento médico em língua portuguesa no século XVIII", são analisados e disponibilizados[19] textos médicos desse período, reunidos em um *corpus* de arquivos de transcrições (Lazzari; Finatto, 2023). Esse *corpus* é de acesso público e oferece manuais escritos por médicos, enfermeiros e cirurgiões. Nele, têm destaque as obras do médico Curvo Semedo (1635-1719) e o primeiro manual de Enfermagem publicado em português, de 1741.

O objetivo desse projeto é descrever e sistematizar terminologias, conceitos e modos de dizer relacionados a doenças e seus tratamentos, criando-se bases para um futuro hiperdicionário (Wives, 1997) de epidemiologia histórica luso-brasileira. Nesse hiperdicionário, a ideia é associar nomes de doenças, partes do corpo, perfis de pessoas/doentes, dados geográficos e datas, com nomes de remédios, de processos e de tratamentos que chegam aos seus correspondentes terapêuticos atuais. Esses dados podem ser correspondidos à identificação das Entidades Nomeadas ao longo dos textos (Zilio; Finatto; Vieira, 2022).

Naturalmente, há todo um trabalho de geração e tratamento desse *corpus* histórico, que é feito também com auxílio de recursos e técnicas de PLN. Isso porque, no período dessas obras, verificava-se toda a sorte de variação de escrita e de tipografia, de modo que em uma mesma página, de um mesmo autor, encontra-se, por exemplo a palavra "água" escrita como "agua" ou "agoa", além de haver itens de vocabulário hoje desconhecido, como o termo anatômico "madre" por "útero". Por isso, o projeto já tenta aproximar, com apoio computacional, a escrita antiga da sua forma atualizada, investindo-se na normalização automática dos textos (Zilio; Lazzari; Finatto, 2024).

Embora sejam enfrentados problemas com a qualidade das informações automaticamente extraídas em dados brutos, pois lida-se com textos apenas na ortografia antiga, plenos de variações nas formas das palavras conforme escritas na época, ferramentas genéricas de PLN já se mostram capazes de identificar informações relevantes (Quaresma; Finatto, 2020). São ferramentas com bom potencial para ajudar especialistas que lidam com esses textos na criação de bases de conhecimento histórico. Enfim, o processamento desse *corpus* ajuda a identificar e refinar a informação nele contida. E isso pode ser muito relevante para

[19] https://sites.google.com/view/projeto38597

várias aplicações, produzindo-se verdadeiros "mapas de navegação" da sua organização e conteúdos.

A ligação de ontologias é, por fim, outro dos objetivos desse projeto de pesquisa. Assim, as antigas terminologias de Saúde poderão ser encontradas e mapeadas em seus equivalentes em ontologias da atualidade (Schriml et al., 2012). Este mapeamento será mais uma forma de enriquecer os recursos derivados do *corpus* e estabelecer quais as áreas da Medicina, da Anatomia e que tipo de doenças e medicamentos eram conhecidos daquela época. Atualmente, as terminologias e expressões relacionadas, conforme usadas nos manuais de Curvo Semedo, já estão contrastadas com o vocabulário de Enfermagem empregado em 1741 (Finatto; Gonçalves; Lazzari, 2023). Esse contraste, entre outros, pode ajudar a identificar diferentes perspectivas, gêneros textuais, conceitos e saberes dos vários práticos e profissionais de Saúde em ação e formação no século XVIII.

27.5 Considerações finais

Há um grande potencial de aplicações de PLN na área de humanidades digitais, muitas ainda a serem exploradas. Como visto neste capítulo, o PLN se faz presente no tratamento de problemas em diferentes fases, desde as etapas iniciais relacionadas à transcrição de manuscritos, ao tratamento de textos como imagem com técnicas como o OCR, a adição de metadados, e a normalização textual. Para as fases de processamento, a partir de um determinado texto 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. O ideal é que exista uma atenção especial aos padrões de disponibilização de dados, como os propostos pelos princípios *FAIR* de Dados Abertos.

Nessa área de intersecção entre humanidades e tecnologias é essencial um trabalho interdisciplinar. São várias questões relevantes a serem combinadas: os objetivos de um pesquisador da área de humanidades, como linguística e literatura, ciências sociais, história, e, na perspectiva do processamento de língua, os objetivos de encontrar, aplicar ou mesmo desenvolver soluções apropriadas para cada tipo de problema, maximizando a correção e a performance das soluções encontradas de acordo com a disponibilidade de recursos.

É crucial que uma perspectiva de IA, centrada no ser humano, seja levada em consideração para fornecer interfaces de usuário adequadas para preparar e acessar as fontes e os dados extraídos. Os projetos, muitas vezes considerando coleções distintas, ou mesmo possuindo objetivos diferentes, podem se beneficiar com a troca de experiências e com o uso das mesmas ferramentas para lidar com os textos e seus conhecimentos codificados.

No contexto do evento principal da língua portuguesa, o PROPOR, tem sido organizado o *Workshop* de Humanidades Digitais e Processamento de Língua Natural, que em 2024 estará na sua terceira edição. Nos anais desse *workshop* disponíveis *on-line*[20], podem ser encontrados mais detalhes sobre os projetos mencionados neste capítulo, e muitos outros.

[20]https://sites.google.com/view/dhandnlp-propor

Agradecimentos

Os trabalhos aqui mencionados contam com o suporte da Fundação de Ciência e Tecnologia (FCT) - projetos CEECIND/01997/2017, 2022.07730.PTDC, UIDB/ 00057/2020 (https://doi.org/10.54499/UIDB/00057/2020), e e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - processos 308926/ 2019-6 (PQ) e 401770/2022-2 (PDE).

Referências

ACKEL, A. Abordagens digitais para estudos de Paleografia: desafios, atualidade, desdobramentos. **LaborHistórico**, v. 7, n. 3, p. 100–120, 2021.

ALBUQUERQUE, G. et al. Applying event classification to reveal the Estado da Índia. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2024.

AREVALO, E. M.; FONTEYN, L. **MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)**. ICON Workshop on Natural Language Processing for Digital Humanities. **Anais**...2021.

BANZA, A. P. **A edição digital da História do Futuro, de António Vieira: arquivo e ferramentas**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

BORIN, E.; DONATO, F. Financial Sustainability of Digitizing Cultural Heritage: The International Platform Europeana. **Journal of Risk and Financial Management**, v. 16, n. 10, p. 421, 2023.

BRITTO, H.; FINGER, M.; GALVES, C. Computational and linguistic aspects of the construction of The Tycho Brahe Parsed Corpus of Historical Portuguese. **Romanistische Korpuslinguistik, Korpora und gesprochene Sprache, Romance Corpus Linguistics, Corpora and Spoken Language, ScriptOralia, v. 126.**, 2002.

CAMERON, H. F.; GONÇALVES, M. F.; QUARESMA, P. **Linguistic and orthographical classic Portuguese variants Challenges for NLP**. Proceedings of the 14th International Conference on the Computational Processing of Portuguese. **Anais**...2020.

CAMERON, H. F.; OLIVAL, F.; VIEIRA, R. Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758). **Revista LaborHistórico**, v. 9, n. 1, p. e52234, 2023.

CLARKE, D. J. B. et al. FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. **Cell Systems**, v. 9, n. 5, p. 417–421, 2019.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 48, p. 88–99, 2017.

COX, A. N. A. M., Simon J. D. AND Gonzalez-Beltran. Ten simple rules for making a vocabulary FAIR. **PLOS Computational Biology**, v. 17, n. 6, p. 1–15, jun. 2021.

DEVARAJU, A. et al. **FAIRsFAIR Data Object Assessment Metrics 0.5**. [s.l.] Research Data Alliance (RDA), out. 2020. Disponível em: <https://zenodo.org/record/6461229>.

FAIR DATA MATURITY MODEL WORKING GROUP RDA. **FAIR Data Maturity**

Model. Specification and Guidelines. Research Data Alliance; Zenodo, 2020. Disponível em: <https://doi.org/10.15497/rda00050>

FINATTO, M. J. B. Humanidades digitais e estudos históricos do léxico. **Domínios de Lingu@gem**, v. 17, p. e1769, 2023.

FINATTO, M. J.; GONÇALVES, M. F.; LAZZARI, R. Léxico e terminologia em um novo gênero textual do século XVIII: o manual para enfermeiros. **In: Natalia Terrón Vinagre & Jenny Brumme (orgs.) Emergencia de nuevos géneros textuales y terminología en la historia de los lenguajes de especialidad.**, 2023.

FINGER, M. Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2000.

GARIJO, D.; POVEDA-VILLALÓN, M. Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web. **CoRR**, v. abs/2003.13084, 2020.

GUIZZARDI, G. Ontology, Ontologies and the "I" of FAIR. **Data Int.**, v. 2, n. 1-2, p. 181–191, 2020.

JACOBSEN, A. et al. **FAIR principles: interpretations and implementation considerations**. **Data intelligence**MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 2020.

KAHLE, P. et al. **Transkribus-a service platform for transcription, recognition and retrieval of historical documents**. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). **Anais**...IEEE, 2017.

KOHO, M. et al. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. **Semantic Web – Interoperability, Usability, Applicability**, v. 12, n. 2, p. 265–278, 2021.

LAZZARI, R. R.; FINATTO, M. J. B. Exame do vocabulário médico no Português no século XVIII: contribuições da lexicometria para o desenho de um dicionário histórico. **Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)**, v. 7, n. 1, p. 102–123, 2023.

LIU, Y.; HEARNE, J.; CONRAD, B. **Recognizing proper names in ur iii texts through supervised learning**. The Twenty-Ninth International Flairs Conference. **Anais**...2016.

LOPEZ-GAZPIO, I. et al. Interpretable semantic textual similarity: Finding and explaining differences between sentences. **Knowledge-Based Systems**, v. 119, p. 186–199, 2017.

NAIR, S. S.; JEEVEN, V. A brief overview of metadata formats. **DESIDOC Journal of Library & Information Technology**, v. 24, n. 4, 2004.

OLIVAL, F.; CAMERON, H.; VIEIRA, R. **As Memórias Paroquiais: do manuscrito ao digital**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

QUARESMA, P.; FINATTO, M. J. B. **Information Extraction from Historical Texts: a Case Study.** DHandNLP@ PROPOR. **Anais**...2020.

RIBEIRO, A. S. **O projecto MONSOON: perspectivas digitais da Índia portuguesa**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

SACRAMENTO, A. DA S. B.; SOUZA, M. **Joint Event Extraction with Contextualized Word Embeddings for the Portuguese Language**. Brazilian Conference on Intelligent Systems. **Anais**...Springer, 2021.

SANTOS, J. et al. Named entity recognition specialised for Portuguese 18th-century History

research. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2024.

SCHRIML, L. M. et al. Disease Ontology: a backbone for disease semantic integration. **Nucleic acids research**, v. 40, n. D1, p. D940–D946, 2012.

SOUSA, M. C. P. DE. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e linguística portuguesa**, v. 16, n. esp., p. 53–93, 2014.

STRIEN, D. VAN et al. **Assessing the impact of OCR quality on downstream NLP tasks**. ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence. **Anais**...2020.

SUN, C.; EMONET, V.; DUMONTIER, M. **A Comprehensive Comparison of Automated FAIRness Evaluation Tools**. (K. Wolstencroft et al., Eds.)13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2022, Virtual Event, Leiden, The Netherlands, January 10th to 14th, 2022. **Anais**...: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <http://ceur-ws.org/Vol-3127/paper-6.pdf>

VIEIRA, R. et al. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. **Journal of Open Humanities Data**, v. 7, p. 20, 2021.

WILKINSON, M.; DUMONTIER, M.; AALBERSBERG, ET AL. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1–9, 2016.

WILKINSON, M.; DUMONTIER, M.; SANSONE, ET AL. Evaluating FAIR maturity through a scalable, automated, community-governed framework. **Sc. Data**, v. 6, n. 1, p. 1–12, 2019.

WIVES, L. K. **Técnicas de Recuperação de Informações Com Ênfase em Informações Textuais**. tese de doutorado—[s.l.] Universidade Federal do Rio Grande do Sul, 1997.

ZILIO, L.; FINATTO, M. J.; VIEIRA, R. **Named Entity Recognition Applied to Portuguese Texts from the XVIII Century**. (C. Trojahn et al., Eds.)Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022. **Anais**...: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <http://ceur-ws.org/Vol-3128/paper10.pdf>

ZILIO, L.; LAZZARI, R. R.; FINATTO, M. J. B. NLP for historical Portuguese: Analysing 18th-century medical texts. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2024.