Infraestrutura de Investigação para a Ciência e Tecnologia da Linguagem PORTULAN CLARIN

António Branco[*], *Amália Mendes*[**] *e Paulo Quaresma*[***]

[*] NLX—Grupo de Fala e Linguagem Natural, Faculdade de Ciências, Universidade de Lisboa,
[**] Centro de Linguística, Faculdade de Letras, Universidade de Lisboa
[***] Departamento de Informática, Escola de Ciências e Tecnologia, Universidade de Évora

Abstract

This paper presents the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language, which is part of the European research infrastructure CLARIN ERIC as its Portuguese national node, and belongs to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance. The PORTULAN CLARIN includes a helpdesk, a repository, where resources, such as corpora, lexicons and processing tools are deposited for long-term archiving and can be searched and retrieved, and a workbench, where Language Technology tools and applications are made readily available online and can be used in different types of interfaces. Its goal is to contribute to the technological development of natural languages and for their preparation for the digital age, with a special focus on the Portuguese language in all its varieties and modalities.

***Keywords*:** research infrastructure, language science, language technology, language resources, Portuguese language

Palavras-chave: infraestrutura para a investigação, ciência da linguagem, tecnologia da linguagem, recursos linguísticos, língua portuguesa

1. Introdução

Este artigo apresenta a Infraestrutura de Investigação para a Ciência e a Tecnologia da Linguagem PORTULAN CLARIN,[1] que constitui o nó nacional da infraestrutura europeia de investigação CLARIN ERIC[2] e faz parte do Roteiro Nacional de Infraestruturas de Investigação de Interesse Estratégico desde a sua versão inicial.[3] Esta infraestrutura tem como objetivo servir a comunidade académica e empresarial que de alguma forma usa a língua portuguesa, ou outras línguas, e permitir o rápido acesso a um repositório que contém um conjunto muito significativo de recursos linguísticos e a uma bancada de trabalho que disponibiliza um vasto leque de serviços de processamento das línguas naturais.

A ampla adesão da comunidade científica a esta infraestrutura, que se verificou desde o início da fase de implementação, permitiu o rápido crescimento da lista de recursos e serviços que integram o repositório da infraestrutura e que estão disponíveis para serem reutilizados. A rede de parceiros de implementação inclui 3 instituições proponentes e mais de 20 centros de investigação em Linguística, Informática, Psicologia, Engenharia, entre outras áreas, em Portugal e no Brasil. O esforço desta vasta rede tem como objetivo a preservação, divulgação, partilha e acessibilidade de recursos linguísticos, incluindo corpora, bases de dados, léxicos, bases de imagens, ferramentas de processamento, entre outros.

[1] https://portulanclarin.net/
[2] https://www.clarin.eu
[3] https://www.fct.pt/media/docs/Portuguese_Roadmap_Infrastructures2020.pdf

A missão da PORTULAN CLARIN é providenciar serviços a todos os utilizadores que de alguma forma estudam ou utilizam a linguagem, nas áreas de Linguística, Inteligência Artificial, Humanidades, Ciência Cognitiva, Psicologia, etc. Os seus objetivos estão totalmente alinhados com as melhores práticas da Ciência Aberta, ao apoiar os utilizadores que pretendem tornar os seus recursos acessíveis a uma comunidade alargada.

A infraestrutura PORTULAN CLARIN segue as características dos vários centros nacionais da infraestrutura europeia CLARIN ERIC (de Jong *et al*., 2020), estabelecendo ligações com os vários polos existentes e com o seu Virtual Language Observatory – VLO (van Uytvank *et al*., 2012), um inventário geral de recursos linguísticos constantes de todos os centros.

Dado o conjunto amplo de recursos disponibilizados, parece-nos importante divulgar a infraestrutura junto da comunidade de investigadores, para que possam utilizar o repositório e a bancada, assim como o balcão de apoio ao utilizador, e para aí poderem conservar e disponibilizar os seus próprios recursos.

Começamos por apresentar a missão da infraestrutura na secção 2, as várias etapas de desenvolvimento da infraestrutura na secção 3, bem como as certificações que a PORTULAN CLARIN recebeu. A secção 4 é dedicada à discussão das características do Repositório, bem como as opções de distribuição e de licenciamento dos recursos. A secção 5 apresenta a bancada de trabalho, a secção 6 o balcão de apoio ao utilizador e o serviço de consultoria e a secção 7 descreve a estrutura de governação e a rede de parceiros. Concluímos na secção 8.

2. Missão

A missão da infraestrutura PORTULAN CLARIN é apoiar investigadores, cientistas, estudantes, profissionais da linguagem e utilizadores em geral cujas atividades dependem de resultados da Ciência e Tecnologia da Linguagem através da distribuição de recursos científicos, do fornecimento de apoio tecnológico, da prestação de consultoria e da disseminação científica.

Nesse sentido, a infraestrutura apoia as atividades em todas as áreas científicas e culturais, com especial relevo para aquelas que estão mais diretamente ligadas à linguagem – quer tendo esta como seu objeto mais imediato, quer como instrumento para abordar os seus tópicos de interesse –, incluindo entre outras, as áreas das Humanidades, Artes e Ciências Sociais, Inteligência Artificial, Ciência Cognitiva e Ciência da Computação, Saúde, Ensino e Promoção da Linguagem, Criatividade Cultural, Herança Cultural, etc. Pretende servir todos aqueles cuja atividade requer a manipulação e exploração de recursos linguísticos, incluindo serviços e dados:

- em todos os tipos de modalidades – língua falada, escrita, gestual, multimodal, etc.;
- em todos os tipos de representações – áudio, texto, vídeo, registo de atividade neurológica, etc.;
- e em todos os tipos de funções – instrumento de comunicação, objeto simbólico, capacidade cognitiva a ser estimulada através de educação formal em língua materna, veículo de conhecimento, capacidade a ser exercitada na aquisição de uma segunda língua, reflexo da atividade mental, forma natural de interação com dispositivos e agentes artificiais, etc.;

Há um vasto leque de áreas e de situações em que o conteúdo da infraestrutura é de extrema utilidade. Não é viável listá-las todas, dada a riqueza dos materiais, sendo previsível que a progressiva utilização por profissionais de áreas diversas venha a revelar várias novas potencialidades. Apresentam-se alguns exemplos que procuram ser ilustrativos dessa diversidade, podendo o utilizador pretender:

- usar uma ferramenta de processamento de linguagem – e.g. conjugadores, extratores de terminologia, concordanciadores, etiquetadores morfossintáticos, gramáticas para processamento linguístico profundo, etc.;
- aceder a conjuntos de dados – e.g. corpora interpretados linguisticamente, bancos terminológicos, registos EEG de experiências neurolinguísticas, coleções de textos literários, etc.;
- obter uma amostra de dados – e.g. enunciados de linguagem gestual de crianças surdas registados em vídeo, palavras para conceitos na subontologia de Organizações, etc.;

- usar aplicações específicas de apoio à investigação – e.g. extratores de frequências de lemas, anotadores de treebanks, etc.;
- usar uma bancada de trabalho online devidamente equipada – para apoiar trabalho de campo na documentação de línguas em risco de extinção, fazer investigação sobre tradução automática, etc.;

Para além da disponibilização de um fácil acesso a um grande conjunto de dados e ferramentas, a PORTULAN CLARIN tem como objetivo contribuir para a preservação do património científico no que diz respeito à língua portuguesa. De facto, os produtores de conteúdos deparam-se frequentemente com dificuldades de curadoria dos recursos, quer por alterações nas equipas de investigação, quer por mudanças de suportes físicos e informáticos. A possibilidade de manter os dados num centro de dados com gestão profissional assegura a perenidade dos dados e a sua conservação.

Apresentando um grande número de recursos acessíveis, a infraestrutura PORTULAN CLARIN não apresenta uma listagem de todos os recursos existentes para a língua portuguesa, uma vez que alguns recursos não podem, por exemplo, por razões de direitos de autor, ser distribuídos à comunidade e exigem por isso um tipo de acesso bastante limitado.

Cabe assinalar também que a PORTULAN CLARIN não tem nenhuma função de prossecução de investigação ou de produção de recursos. Agrega os conteúdos que os parceiros de implementação disponibilizam à comunidade mas não produz recursos.

Figura 1: Página de entrada da infraestrutura PORTULAN CLARIN

3. Criação da infraestrutura e certificações internacionais

Em 2008, teve início o projeto europeu preparatório para a CLARIN, financiado pela Comissão Europeia: a equipa da Universidade de Lisboa, coordenada por António Branco, foi o parceiro que representou Portugal no consórcio. O projeto foi concluído com sucesso em 2011, sendo a rede para a língua portuguesa, com 18 membros, a maior em toda a CLARIN europeia. Em 2013, a candidatura da PORTULAN CLARIN foi submetida ao concurso para o Roteiro Nacional das Infraestruturas de Investigação por 3 parceiros proponentes e 19 parceiros de implementação, incluindo o Camões, IP, o organismo oficial para a promoção da língua portuguesa. A candidatura foi aprovada em 2014 com a pontuação máxima da comissão de avaliação e a PORTULAN CLARIN tornou-se um membro fundador do primeiro Roteiro Nacional das Infraestruturas de

Investigação de Interesse Estratégico (RNIIE). Ainda em 2014, Portugal torna-se o 11º membro do consórcio europeu para infraestrutura CLARIN ERIC após o seu pedido de adesão ter sido aceite pela Assembleia Geral deste consórcio. Finalmente, em março 2017 a candidatura da infraestrutura PORTULAN CLARIN à chamada de propostas de financiamento de Infraestruturas é aprovada e o projeto decorre de 2017 a 2021.

Os trabalhos de recolha, curadoria e distribuição de recursos linguísticos têm início em 2017 e a primeira versão do website da infraestrutura é lançada em 2019, com o repositório e a bancada já funcionais. Em março 2020 a classificação de "Alto" é atribuída ao estado de maturidade da PORTULAN CLARIN na ronda nacional de avaliação das infraestruturas do RNIIE, pela comissão externa independente designada pela FCT. A infraestrutura recebeu entretanto várias certificações internacionais: o repositório obteve a certificação internacional CoreTrustSeal[4] em 2019;[5] também em 2019, a infraestrutura obteve da CLARIN ERIC a certificação como um centro nacional fornecedor de serviços[6] e a certificação como um centro do conhecimento.[7]

Durante a fase de preparação e de implementação da infraestrutura nacional foram realizados dois workshops, com 10 anos de intervalo. Em março 2010, teve lugar o workshop financiado pela FCT para formar a rede nacional CLARIN. Este decorreu em Lisboa e teve já a participação de 17 unidades de investigação acreditadas pela FCT, um núcleo que se manteve ao longo da fase de implementação e que voltou a reunir-se, com adição de novos parceiros, no workshop da rede de parceiros de implementação, em dezembro de 2020.

De 2008 a 2021, a PORTULAN CLARIN criou uma alargada rede de parceiros em Portugal e no Brasil. Esta rede, que esteve envolvida no projeto europeu preparatório (2008-2011) e, depois, no projeto de implementação (2018-2021), colaborou ativamente na criação da infraestrutura online com um leque de recursos e de serviços muito abrangente, e que continuará a crescer.

4. Repositório

O repositório constitui uma componente fundamental na infraestrutura PORTULAN CLARIN. Os recursos disponibilizados pelos respetivos proprietários passam a figurar no catálogo de recursos disponíveis no repositório de onde podem ser acedidos por outros utilizadores, mediante as condições de licenciamento estipuladas pelos detentores dos direitos.

Explicita-se primeiramente o procedimento para o depósito de recursos no repositório e de seguida o processo de pesquisa e acesso aos recursos constantes do repositório.

4.1 Depositar um recurso

Para depositar um recurso, os utilizadores autenticados na plataforma preenchem um formulário de depósito de recurso, com metadados, que poderão sempre ser atualizados mais tarde. Os metadados são fundamentais para os utilizadores da plataforma pois são as informações sobre a tipologia do recurso que possibilitam depois a pesquisa por tipo, língua, modalidade, entre outros campos. Durante o depósito do recurso, a equipa de gestão acompanha o depositante e dá apoio, se necessário, ao preenchimento dos metadados, verificando a sua integridade e boa formação.

O depositante deve ainda conceder à infraestrutura PORTULAN CLARIN uma licença de depósito e distribuição, não exclusiva, isto é, o utilizador é livre de distribuir ou publicar o recurso noutras plataformas e infraestruturas. Esta licença é apenas para distribuição e, portanto, não transfere a propriedade ou direitos de

[4] https://www.coretrustseal.org/
[5] https://www.coretrustseal.org/wp-content/uploads/2019/12/PORTULAN-CLARIN.pdf
[6] https://www.clarin.eu/content/clarin-centres
[7] https://www.clarin.eu/content/knowledge-centres

autoria à infraestrutura, tendo aliás os metadados campos específicos para indicação da autoria e da propriedade do recurso.

Na fase seguinte, e após o envio dos dados, a equipa de gestão confere a correspondência entre os dados a depositar e a sua descrição nos metadados e atribui um identificador persistente a cada recurso. No final do processo, a entrada relativa ao recurso e aos seus metadados fica publicamente acessível no repositório online.

4.2 Associar uma licença de utilização a um recurso depositado

A PORTULAN CLARIN adere às políticas de Ciência Aberta, Acesso Aberto, Dados Abertos e Código Aberto, mas, com vista a maximizar o benefício para os seus utilizadores, deixa total liberdade aos distribuidores de recursos para escolherem a licença de distribuição que entenderem ser a mais adequada. Esta abordagem permite que a infraestrutura divulgue um largo conjunto de recursos, que podem ter formas de acesso diferenciadas, e que os autores dos recursos sintam que a infraestrutura se adequa às especificidades várias dos dados com que trabalham. Os utilizadores que depositam recursos têm à sua escolha um vasto conjunto de licenças, quer licenças sem qualquer restrição de acesso, quer licenças de uso restrito (por exemplo, apenas para uso não comercial, etc.), que requerem a identificação dos utilizadores, de acordo com os termos exigidos pelo depositante. Para além de licenças amplamente divulgadas, como as Creative Commons, há ainda licenças META-SHARE e licenças da European Language Resources Association (ELRA). Caso o depositante precise de ajuda para identificar a licença mais adequada ao seu recurso, pode contactar o balcão de apoio aos utilizadores da PORTULAN CLARIN e receber informações sobre as diferentes possibilidades e seus efeitos práticos no acesso dos utilizadores da infraestrutura ao recurso.

O repositório da PORTULAN CLARIN fornece armazenamento e distribuição de dados. A responsabilidade de seguir as normas legais e éticas para armazenamento e distribuição de dados é do repositório. A responsabilidade de seguir as normas legais e éticas para a criação e compilação de dados é do depositante dos dados. Conforme foi mencionado, para depositar um recurso na PORTULAN CLARIN, o depositante deve preencher e apresentar um acordo de licenciamento de depósito e distribuição. Nesse acordo, fica explicitamente reconhecido pelo depositante que as normas legais e éticas foram cumpridas no momento da criação do recurso. O depositante também deve especificar se o recurso contém dados com algum grau de confidencialidade e a presença de tais dados restringirá o conjunto de possíveis licenças e utilizadores finais.

4.3 Pesquisa de recursos no repositório

Os recursos guardados e distribuídos através do repositório são listados na página de entrada do repositório. A ordem de apresentação pode ser alterada, existindo as opções de ordenação por nome do recurso, língua, tipo de recurso, tipo de modalidade, tendo todas as opções a possibilidade de ordenação alfabética crescente ou decrescente. Além disso, é possível pesquisar por palavra-chave, que recupera entradas do repositório que contenham essa palavra no nome do recurso ou na sua descrição. Esta pesquisa incide sobre os metadados do recurso e não sobre os dados linguísticos contidos no recurso. Finalmente, uma terceira opção para localizar um recurso é usar o conjunto de filtros que são apresentados à direita da página. Estão disponíveis, neste momento, os seguintes filtros: língua, variedade, período temporal, área geográfica, modalidade, número de línguas incluídas, natureza dos dados multilingues (comparáveis, paralelos, etc.), domínio, tópico, tipo de aplicação do recurso, tipo de recurso, media, formato. Os filtros podem ser combinados. Por exemplo, se no atual estado do repositório, o utilizador procurar um corpus oral de português, poderá começar por usar o filtro "Resource type=corpus", que restringirá a lista a 266 recursos, aplicar de seguida o filtro "Language=Portuguese", que reduz a 102 recursos e depois o filtro "Modality type=spoken language", obtendo assim 6 resultados que correspondem à sua pesquisa.

Este conjunto de filtros resulta diretamente do preenchimento dos metadados por parte do depositante e apenas estão disponíveis os campos para os quais existe algum recurso disponível. Assim, se um novo tipo de recurso for depositado e descrito nos metadados, esse campo passará a estar automaticamente visível para ser selecionado nas opções de filtro dos recursos.

Periodicamente, os registos de metadados são automaticamente recolhidos no Virtual Language Observatory (VLO),[8] que atua como um hub de pesquisa central para todo o ecossistema de repositórios CLARIN pan-europeu.

A pesquisa do conteúdo dos dados, por sua vez, é possível por meio da funcionalidade Federated Content Search (FCS) do CLARIN.[9] Essa funcionalidade permite, a partir de uma única interface online, executar uma consulta por uma expressão em vários conjuntos de dados, que são distribuídos em diferentes nós CLARIN nacionais e aí se encontram depositados.

A Figura 2 (a) mostra a página de busca do repositório da PORTULAN CLARIN, com a lista de recursos ordenados alfabeticamente. A caixa de texto na parte superior é usada para pesquisa por palavra-chave, enquanto as opções da direita permitem filtrar os resultados por vários critérios. A Figura 2 (b) mostra o exemplo de uma página de um recurso.

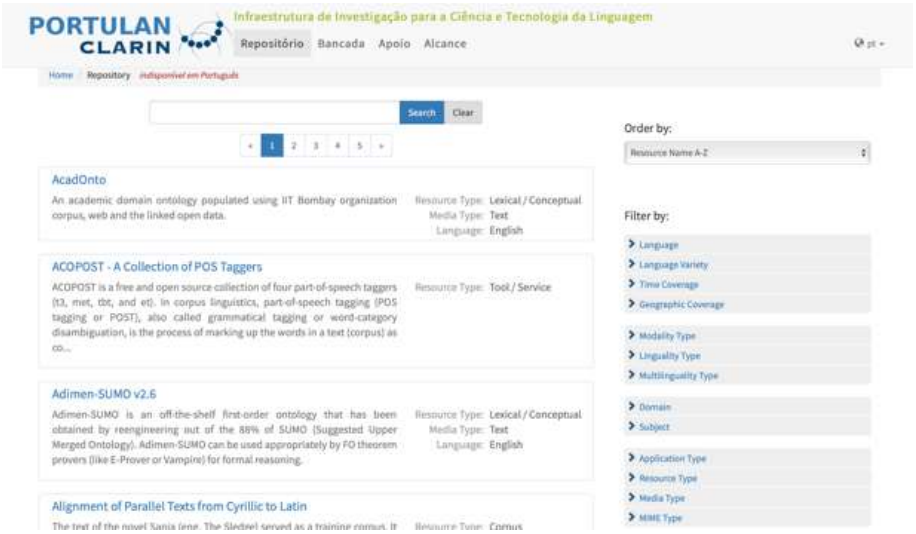

Figura 2 (a): Repositório da PORTULAN CLARIN

[8] https://vlo.clarin.eu/
[9] https://contentsearch.clarin.eu/

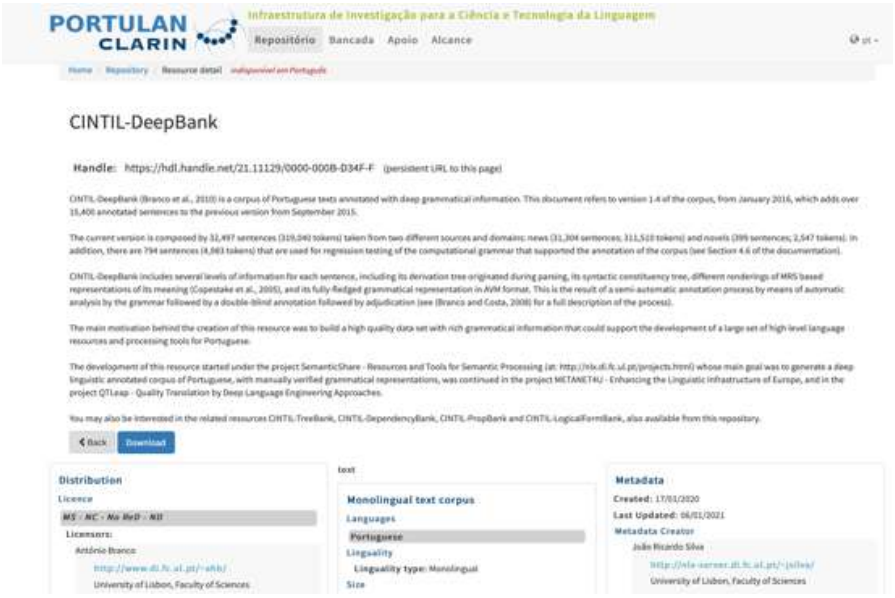

Figura 2 (b) Página de um recurso no repositório

Após identificar o recurso e entrar na respetiva página, o utilizador pode aceitar a licença que lhe estiver associada e descarregar o recurso (opção "download"), como mostra a Figura 3, ou é reencaminhado para a pessoa de contacto do recurso, caso a licença tenha algum tipo de restrição.

Figura 3: Exemplo de licença associada a um recurso

4.4 Aspetos técnicos do repositório

O repositório é construído com base na plataforma Django.[10] O esquema de banco de dados subjacente e a lógica do fluxo de trabalho foram desenvolvidos adaptando e melhorando o software do repositório METASHARE[11] anteriormente disponível. O site do repositório é criado com a estrutura de front-end Bootstrap[12] CSS, que fornece uma interface consistente que permite acessos por computador pessoal e por plataforma móvel. A funcionalidade de pesquisa por palavra-chave é feita através do Apache Solr.[13] A recolha automática de metadados para o hub de pesquisa central VLO usa o protocolo OAI-PMH[14] para interoperabilidade de repositórios.

5. Bancada de trabalho

Outra componente importante da PORTULAN CLARIN é a operação online de ferramentas e aplicações de Tecnologia da Linguagem. Isso é realizado por meio de uma bancada de trabalho que disponibiliza uma ampla gama de serviços de processamento e aplicações. Contrariamente ao repositório, que permite, caso a licença assim autorize, descarregar dados ou ferramentas que serão utilizados nos computadores pessoais do utilizador, a bancada de trabalho fornece serviços para serem usados diretamente a partir de interfaces online da plataforma. A lista dos tipos de serviços disponíveis na bancada, ou em preparação, no momento da escrita do presente artigo encontra-se no Anexo A.[15] Tal como acontece com o repositório, também a bancada de trabalho continuará a integrar mais tipos de serviços de processamento, e para cada tipo mais ferramentas diferentes.

Os serviços são apresentados na página de entrada da bancada por categorias (por exemplo, anotação morfossintática, reconhecimento de entidades nomeadas, análise de sentimentos, etc.). A Figura 4 ilustra a página principal da bancada de trabalho. Como é visível, cada tipo de serviço pode estar associado a mais de uma ferramenta com que é possível trabalhar.

5.1 Formas de utilização

Os serviços e aplicações são disponibilizadas em diferentes modos de interação na bancada da PORTULAN CLARIN. Um dos modos de interação é o uso das ferramentas por meio do browser: neste caso, o utilizador escreve ou copia uma frase, ou mesmo um texto um pouco mais extenso, numa caixa de texto, dá ordem para processar esses dados e obtém o resultado imediatamente. A Figura 5 mostra um exemplo desse modo de interação.

[10] https://www.djangoproject.com/
[11] http://www.meta-share.org/
[12] https://getbootstrap.com/
[13] https://lucene.apache.org/solr/
[14] https://www.openarchives.org/pmh/
[15] A bancada de trabalho da PORTULAN CLARIN é composta por um conjunto de serviços de processamento da linguagem que assentam num vasto conjunto de pesquisas de diferentes autores e equipas, que são aqui reconhecidas: Barreto *et al.* (2006), Branco *et al.* (2003, 2006, 2010, 2011, 2012, 2014), Costa *et al*. (2012), Cruz *et al*. (2018), Miranda *et al*. (2011), Rodrigues *et al.* (2016, 2020), Santos *et al.* (2019), Silva *et al*. (2009), Veiga *et al.* (2011).

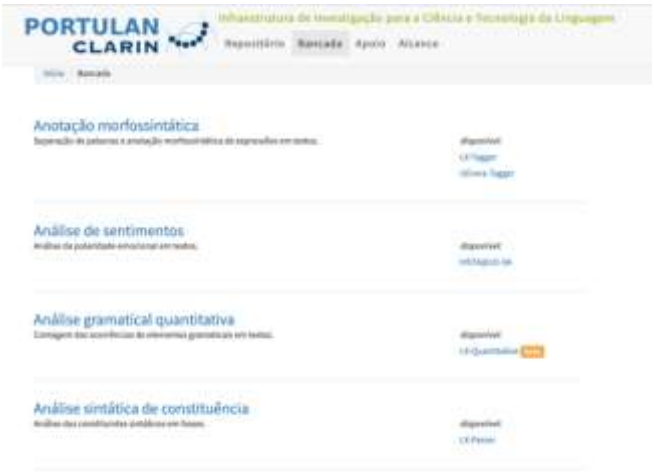

Figura 4: Página da Bancada

Figura 5: Exemplo de serviço na Bancada

Embora a interação direta por meio do browser seja útil para pequenas quantidades de dados, ou como uma demonstração das capacidades e formato de saída de uma ferramenta, grandes quantidades de dados requerem outros modos de interação. Para certas ferramentas, a bancada também permite o carregamento de ficheiros e arquivos para processamento (opção "processar ficheiros"). A Figura 6 mostra a janela que abre após a escolha desta opção. A tarefa de processamento dos arquivos carregados será internamente adicionada a uma fila de espera e tratada de forma assíncrona. Após a conclusão do processamento dos arquivos, o utilizador será notificado por correio eletrónico e poderá descarregar o resultado a partir de um URL exclusivo, gerado quando a tarefa foi realizada. Os ficheiros processados são de seguida apagados do servidor, não sendo guardada cópia dos dados submetidos e processados.

Um terceiro modo de interação permite aceder às ferramentas como serviços da web. Isso é particularmente útil para utilizadores finais que desejam integrar a funcionalidade e os resultados das

ferramentas de processamento no seu próprio código, que estão a programar, qualquer que seja a linguagem de programação que estiverem a usar, sem terem que se preocupar em instalar a ferramenta localmente nas suas próprias máquinas. É também útil para depositantes que desejam tornar a funcionalidade da ferramenta acessível sem disponibilizar a própria ferramenta. As ferramentas disponibilizadas como um serviço da web têm uma interface que pode ser facilmente chamada remotamente.

Figura 6: Carregamento de ficheiro para processamento

5.2 Aspetos técnicos da bancada

As ferramentas e aplicações podem ter sido implementadas pelos seus autores com diferentes linguagens (C, Java, Perl, Python, etc.) e podem necessitar de diferentes bibliotecas de software de suporte. Para lidar de forma mais adequada com esse ambiente heterogéneo, todas as ferramentas são organizadas em containers Docker separados. Isso facilita a sua configuração, minimiza as dependências de todo o sistema e, ao empregar várias instâncias de um container, permite gerir diretamente e de forma equilibrada os fluxos de processamento. A comunicação entre as ferramentas é realizada pelo protocolo padrão XML-RPC. O mesmo protocolo é usado para os serviços da web. O site do workbench também é criado com o Bootstrap, fornecendo uma interface consistente em toda a PORTULAN CLARIN.

6. Apoio e Consultoria

Para além de distribuir e fornecer acesso a recursos linguísticos e serviços de processamento, um dos componentes importantes da infraestrutura PORTULAN CLARIN é dar apoio à comunidade de utilizadores de recursos linguísticos e de tecnologia da linguagem através de um balcão de apoio para os utilizadores e de um serviço de consultoria em Tecnologia da Linguagem para a comunidade em geral.

6.1 Balcão de apoio

A equipa da PORTULAN CLARIN mantém um balcão de apoio ao utilizador para a infraestrutura, para os conjuntos de dados do seu repositório e para as ferramentas de processamento e serviços online que disponibiliza na sua bancada de trabalho. O serviço pode ser solicitado por todos os utilizadores, mas será particularmente útil no caso de estudantes ou de investigadores em início de carreira, ou para utilizadores de áreas de investigação menos familiarizados com os conceitos de recurso linguístico e tecnologia da linguagem.

Além de fornecer ajuda sobre como obter e usar os recursos científicos na infraestrutura e dar apoio na resolução de problemas, o apoio ao utilizador também envolve a curadoria dos recursos enviados. Isso permite, por exemplo, fornecer ajuda na conversão dos recursos depositados para formatos diferentes dos formatos originais, incluindo formatos padrão, que devem ser particularmente úteis para utilizadores que não possuem o conhecimento técnico para fazer a conversão do formato.

6.2 Consultoria

Outro objetivo da CLARIN é a partilha de conhecimento entre os vários centros da infraestrutura europeia para que todos os utilizadores possam beneficiar dos desenvolvimentos conseguidos nos vários nós nacionais. Para tal, foram estabelecidos Centros de Conhecimento (K-centers), que são entidades certificadas centralmente pela CLARIN como tendo capacidade de fornecer consultoria especializada nalguma área.

A PORTULAN CLARIN é reconhecida como um K-Center especializado em Ciência e Tecnologia da língua portuguesa, abordando todos os temas relativos a esta língua: da Fonética ao Discurso e ao Diálogo; considerando todas as funções da linguagem, do desempenho comunicativo à expressão cultural; incluindo todas as disciplinas, da Linguística Teórica à Tecnologia da Linguagem; cobrindo todas as variedades da língua, desde variedades padrão nacionais a dialetos e terminologias; e levando em consideração todos os meios de representação, de gravações de áudio a imagens de atividade cerebral.

7. Governação

7.1 Rede de parceiros

A implementação da infraestrutura está a ser realizada pela execução de um projeto apoiado a nível nacional no âmbito de um concurso aberto e competitivo para propostas de financiamento de projetos de implementação de infraestrutura de investigação do Roteiro RNIIE para todos os domínios científicos.[16] Este projeto conta com três instituições proponentes, a Faculdade de Ciências da Universidade de Lisboa (coordenação), a Faculdade de Letras da Universidade de Lisboa e a Universidade de Évora, encontrando-se a fase de implementação a ser finalizada no momento da escrita do presente artigo. O projeto contou ainda com a contribuição de uma ampla rede de parceiros de implementação, que engloba mais de vinte centros de investigação e instituições, que cobrem uma vasta gama de domínios científicos servidos pela infraestrutura. São parceiros de todas as regiões de Portugal, incluindo as ilhas dos Açores, e ainda do Brasil. O Camões, IP, instituição portuguesa responsável pela política da língua portuguesa, faz parte da rede e contribui para a prossecução da missão da infraestrutura de promoção da língua portuguesa.

Os parceiros encontram-se ativamente envolvidos no depósito de recursos científicos e na disponibilização de serviços na bancada da infraestrutura, e, com a transição do projeto de implementação para a fase de operação da infraestrutura, passaram a integrar o Fórum de Aconselhamento Científico, apresentado no Anexo B. A lista de centros está aberta a novas contribuições, podendo incluir instituições de todos os domínios, incluindo de Humanidades, Inteligência Artificial, Neurociência, Ciência Cognitiva, etc.

[16] Projeto PORTULAN CLARIN – Infraestrutura de Investigação para a Ciência e Tecnologia da Linguagem, com verbas do programa Lisboa2020, Alentejo2020 e da FCT – Fundação para a Ciência e a Tecnologia, com a referência PINFRA/22117/2016.

7.2 Órgãos de governação

A governação da infraestrutura inclui uma Comissão Diretiva e uma Equipa de Gestão, uma Comissão de Aconselhamento Estratégico e um Fórum de Aconselhamento Científico, cuja composição no momento da escrita do presente artigo se encontra apresentada no Anexo B.

8. Conclusão

Este artigo apresenta a Infraestrutura de Investigação para a Ciência e Tecnologia da Linguagem PORTULAN CLARIN, que é o nó nacional português da infraestrutura pan-europeia CLARIN ERIC, com 20 países membros, e faz parte do RNIIE. A sua missão é apoiar o maior número e a mais ampla gama de utilizadores que precisam de recorrer a resultados da Ciência e Tecnologia da Linguagem. Isso é realizado por meio de três pilares principais na infraestrutura: (i) um repositório para o depósito de longo prazo e acesso a recursos linguísticos, sejam eles dados linguísticos ou ferramentas; (ii) uma bancada de Tecnologia da Linguagem que disponibiliza um amplo conjunto de serviços e aplicações de processamento de linguagem, por meio de vários modos de interação online; e (iii) um balcão de apoio e serviços de consultoria que dão ajuda aos seus utilizadores. A infraestrutura segue os princípios da Ciência Aberta e os seus serviços são abertos a todos os que a queiram utilizar, sem necessidade de registo pessoal ou outras barreiras de acesso dispensáveis.

Agradecimentos

Os resultados aqui apresentados foram financiados em parte pelo projeto PORTULAN CLARIN – Infraestrutura de Investigação para a Ciência e Tecnologia da Linguagem, com verbas do programa Lisboa2020, Alentejo2020 e da FCT – Fundação para a Ciência e a Tecnologia, com a referência PINFRA/22117/2016, e também FCT – Fundação para a Ciência e a Tecnologia, com a referência UIDB/00214/2020.

Referências

De Jong, Franciska, Bente Maegaard, Darja Fišer, Dieter Van Uytvanck & Andreas Witt (2020) Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC 2020)*, ELRA.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes & João Silva (2006) Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 1438-1443.

Branco, António, Sérgio Castro, João Silva & Francisco Costa (2011) *CINTILDepBank handbook: Design options for the representation of grammatical dependencies*. Technical Report DI-FCUL-TR-2011-03, Universidade de Lisboa.

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto & João Graça (2010) Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *Proceedings of the 7th International Conference on Language Resourcesand Evaluation (LREC)*, pp. 1810-1815.

Branco, António & Tiago Henriques (2003) Aspects of verbal inflection and lemmatization: Generalizations and algorithms. In *Proceedings of XVIII Annual Meeting of the Portuguese Association of Linguistics (APL)*, pp. 201-210.

Branco, António & Filipe Nunes (2012) Verb analysis in a highly inflective language with an MFF algorithm. In *Proceedings of the 11th International Conference on the Computational Processing of Portuguese (PROPOR)*, 7243 in Lecture Notes in Artificial Intelligence, pp. 1-11.

Branco, António, João Rodrigues, João Silva, Francisco Costa & Rui Vaz (2014) Assessing automatic text classification for interactive language learning. In *Proceedings of the IEEE International Conference on Information Society (iSociety)*, pp. 72-80.

Branco, António & João Silva (2006) A suite of shallow processing tools for Portuguese: LX-Suite. InProceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 179-182.

Costa, Francisco & António Branco (2012) Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 266-275.

Cruz, André Ferreira, Gil Rocha & Henrique Lopes Cardoso (2018) Exploring spanish corpora for portuguese coreference resolution. In *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 290-295.

Miranda, Nuno, Ricardo Raminhos, Pedro Seabra, João Sequeira, Teresa Gonçalves & Paulo Quaresma (2011) Named entity recognition using machine learning techniques. In *EPIA-11, 15th Portuguese Conference on Artificial Intelligence*, pp. 818-831.

Rodrigues, João, António Branco, Steven Neale & João Silva (2016) LX-DSemVectors: Distributional semantics models for the Portuguese language. In *Proceedings of the 12th International Conference on the ComputationalProcessing of Portuguese (PROPOR'16)*, pp. 259-270.

Rodrigues, João, Francisco Costa, João Silva & António Branco (2020) Automatic syllabification of portuguese. *Revista da Associação Portuguesa de Linguística*, pp. 715-720.

Santos, Rodrigo, João Silva, António Branco & Deyi Xiong (2019) The direct path may not be the best: Portuguese-chinese neural machine translation. In *Proceedings of the 19th EPIA Conference on Artificial Intelligence*, pp. 757-768.

Silva, João, António Branco, Sérgio Castro & Ruben Reis (2009) Out-of-the-box robust parsing of Portuguese. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 75-85.

Van Uytvanck, Dieter, Herman Stehouwer & Lari Lampen (2012) Semantic metadata mapping in practice: the Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, ELRA, pp. 1029-1034.

Veiga, Arlindo, Sara Candeias & Fernando Perdigão (2011) Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment. In *Proceedings of the 8th Brazilian Symposiumin Information and Human Language Technology*.

Anexo A Catálogo de serviços de processamento na bancada de trabalho

Anotação morfossintática
Separação de palavras e anotação morfossintática de expressões em textos.

Análise de sentimentos
Análise da polaridade emocional em textos.

Análise gramatical quantitativa
Contagem das ocorrências de elementos gramaticais em textos.

Análise sintática de constituência
Análise dos constituintes sintáticos em frases.

Análise sintática de dependências
Análise das funções gramaticais em frases.

Análise sub-sintática
Separação de palavras, lematização, análise flexional e anotação morfossintática de expressões em textos.

Análise temporal
Análise de eventos e de informação temporal em textos.

Busca em treebank
Seleção de padrões sintáticos e de expressões em frases anotadas.

Classificação de proficiência
Análise quantitativa e classificação do nível de proficiência de textos.

Concordanceamento
Seleção de contextos de ocorrência em textos anotados.

Conjugação verbal
Conjugação de verbos.

Consulta da wordnet
Consulta da rede de semântica lexical wordnet.
(em preparação)

Etiquetador de papéis semânticos
Análise de papéis semânticos de constituintes sintáticos em frases.

Extração de palavras chave
Extração de palavras chave em textos.

Flexão nominal
Lematização e flexão de expressões nominais.

Lematização verbal
Lematização e flexão de expressões verbais.

Normalização ortográfica
Conversão para norma ortográfica.
(em preparação)

Reconhecimento de entidades nomeadas
Deteção e classificação semântica de nomes em textos.

Resolução de correferência
Para cada entidade referida num texto, identificar nesse texto todas as expressões que a referem.

Semelhança semântica
Semelhança semântica entre palavras.

Separação de frases
Segmentação de textos em frases e parágrafos.

Separação de unidades lexicais
Segmentação de textos em unidades lexicais.

Silabificação
Silabificação de expressões.

Tradução
Tradução de uma frase da língua origem para uma língua destino.

Transcrição
Representação escrita de fala.
(em preparação)

Transcrição fonológica
Conversão de representação grafémica para representação fonológica.

Anexo B Governação

1. Estrutura de gestão

Comissão Diretiva
- Diretor Geral
  - António Branco
- Diretora Executiva
  - Amália Mendes
- Diretor Executivo
  - Paulo Quaresma (até fev. 2021) / Teresa Gonçalves (desde fev. 2021)

Equipa de Gestão
Gestor Técnico
Luís Gomes
Gestor dos Recursos Científicos e Apoio ao Utilizador
João Ricardo Silva
Gestora Administrativa e de Comunicação
Andrea Teixeira

2. Comissão de Aconselhamento Estratégico

Ana Paula Laborinho
Diretora da Organização de Estados Ibero-Americanos para a Educação, a Ciência e a Cultura; ex-Presidente do Instituto Camões (2010-2012) e do Camões, Instituto da Cooperação e da Língua (2012-2017). Professora Universitária e Gestora Cultural.

Daniela Braga
Fundadora e Presidente Executiva da DefinedCrowd, start up dedicada à tecnologia da linguagem e a soluções e serviços para a Inteligência Artificial. Empreendedora.

Nicolau Santos
Presidente do Conselho de Administração da Agência Lusa; ex-Diretor-adjunto do semanário "Expresso" (2006-2018). Jornalista.

3. Fórum de Aconselhamento Científico

Cristina Martins e Margarita Correia, Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Faculdade de Letras Universidade de Coimbra

Pilar Barbosa e Cristina Flores, Centro de Estudos Humanísticos (CEHUM), Universidade do Minho, Portugal

Augusto Silva, Centro de Estudos Filosoficos e Humanísticos, Faculdade de Filosofia, Universidade Católica de Braga, Portugal

José Augusto Leitão, Centro de Investigação do Núcleo para os Estudos e Intervenção Cognitivo-Comportamental (CINEICC), Faculdade de Psicologia, Universidade de Coimbra, Portugal

Amália Mendes, Centro de Linguística da Universidade de Lisboa (CLUL), Faculdade de Letras, Universidade de Lisboa, Portugal

Fátima Oliveira, João Veloso e Rui Silva, Centro de Linguística da Universidade do Porto (CLUP), Faculdade de Letras, Universidade do Porto, Portugal

Maria do Céu Caetano e Raquel Amaro (e Francisca Xavier numa fase anterior), Centro de Linguística da Universidade Nova de Lisboa (CLUNL), Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Portugal

Luís Gomes, Centro ALGORITMI, Universidade dos Açores, Portugal

São Luís Castro, Centro de Psicologia da Universidade do Porto (CPUP), Faculdade de Psicologia e Ciências da Educação, Universidade do Porto, Portugal

António Branco, Faculdade de Ciências, Universidade de Lisboa, Portugal

Paulo Quaresma e Teresa Gonçalves, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal

Nuno Mamede, Instituto de Engenharia de Sistemas e Computadores (INESC), Instituto Superior Tecnico,´ Universidade de Lisboa, Portugal

Ricardo Campos, INESC TEC, Laboratório de Inteligência Artificial e Apoio à Decisão (INESC TEC/LIAAD), Centro de Investigação em Cidades Inteligentes (Ci2 – IPT), Instituto Politécnico de Tomar, Portugal

Fernando Perdigão, Instituto de Telecomunicações Coimbra (IT Coimbra), Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal

Gabriel Lopes e Nuno Marques, Laboratório de Ciência da Computação e Informática (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Eugénio Oliveira e Henrique Lopes Cardoso, Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal

Renata Vieira (e Vera Strube de Lima numa fase anterior), Faculdade de Informática (FACIN), Pontifícia Universidade Católica do Rio Grande do Sul, Brasil

Aline Villavicencio, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brasil

Thiago Pardo, Núcleo Interinstitucional para a Linguística Computacional (NILC), Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, Brasil

Rui Vaz, Camões, I.P. – Instituto da Cooperação e da Língua, Portugal