Infrastructure for the Science and Technology of Language PORTULAN CLARIN

António Branco,[1] **Amália Mendes,**[2] **Paulo Quaresma,**[3]
Luís Gomes,[1] **João Silva,**[1] **Andrea Teixeira**[1]
[1]*University of Lisbon*
NLX—Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
[2]*University of Lisbon*
Center of Linguistics, School of Arts and Humanities
[3]*University of Évora*
Escola de Tecnologia

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. It encompasses a repository, where resources and metadata are deposited for long-term archiving and access, and a workbench, where Language Technology tools and applications are made available through different modes of interaction, among many other services. It is an asset of the utmost importance for the technological development of natural languages and for their preparation for the digital age, contributing to ensure the citizenship of their speakers in the information society.

Keywords: research infrastructure, language science, language technology

1. Introduction

This paper presents the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language,[1] which is part of the European research infrastructure CLARIN ERIC[2] as its Portuguese national node, and belongs to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance.[3] It ensures the preservation and fostering of the scientific heritage regarding natural languages, supporting the preservation, promotion, distribution, sharing and reuse of language resources, including text collections, lexicons, processing tools, etc.

PORTULAN CLARIN includes a repository of language resources and tools, as well as a workbench with language processing services. The expanding list of resources and services results largely from a wide network of implementation partners, formed by 3 proponent partners and over 20 research centers working in Computer Science, Linguistics, Psychology, etc., from Portugal and Brazil.

The mission of PORTULAN CLARIN is to provide services to all kinds of users that in one way or another have to handle or process language, which naturally includes researchers from Artificial Intelligence, Humanities, Cognitive Science, etc.

PORTULAN CLARIN fosters Open Science practices by supporting its users in making their results and resources accessible to all sectors of an inquiring society.

It represents an asset of the utmost importance for the technological development of natural languages and to their preparation for the digital age, contributing to ensure the citizenship of their speakers in the information society.

In this paper we present the goals, target users, and mission of the infrastructure in Section 2. The repository and the workbench are described in Sections 3 and 4, respectively. Section 5 introduces the organization of the Help Desk and Consultancy support to the users of the infrastructure, and Section 6 the sharing and licensing options offered by the platform. Finally, we present in Section 7 the certifications received by PORTULAN CLARIN and in Section 8 its system of governance and network of implementation partners, before offering the concluding remarks in Section 9.

2. Mission

2.1. All about human natural languages

The mission of PORTULAN CLARIN is to support researchers, innovators, citizen scientists, students, language professionals and users in general whose activities resort to research results from the Science and Technology of Language. This is pursued by means of the distribution of scientific resources, the supplying of technological support, the provision of consultancy, and the fostering of scientific dissemination.

2.2. All scientific and cultural domains served

This infrastructure supports activities in all scientific and cultural domains with special relevance to those that are more directly concerned with language—whether as their immediate subject, or as an instrumental mean to address their topics. This includes, among others, the areas of Artificial Intelligence, Computation and Cognitive Sciences, Humanities, Arts and Social Sciences, Healthcare, Lan-

[1] https://portulanclarin.net/
[2] https://www.clarin.eu/
[3] https://www.fct.pt/apoios/equipamento/roteiro/index.phtml.en

Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 1–7
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Figure 1: Front page of the PORTULAN CLARIN research infrastructure

guage Teaching and Promotion, Cultural Creativity, Cultural Heritage, etc.

2.3. All results from research on language shared

The infrastructure serves all those whose activity requires the handling and exploration of language resources, including language data and services:

- in all sorts of modalities—spoken, written, sign, multimodal, etc.
- in all types of representations—audio, text, video, records of brain activity, etc.
- and in all types of functions—instrument for communication, symbolic object, cognitive ability to be stimulated through formal education in native language, knowledge vehicle, ability to be exercised in the acquisition of a second language, reflection of mental activity, natural form of interaction with artificial agents and devices, etc.

It is used when it is necessary, for example:

- to use a language processing tool—e.g. conjugators, terminology extractors, concordancers, part-of-speech taggers, parsers, named entity recognizers, deep linguistic processing grammars, etc.
- to access data sets—e.g. linguistically interpreted corpora, terminology data bases, EEG records of neurolinguistic experiments, transcriptions, collections of literary texts, etc.
- to obtain a data sample—e.g. video recording of deaf children sign language, words for concepts in the Organization subontology, etc.
- to use specific research support applications—e.g. lemma frequency extractors, treebank annotators, etc.
- to use an appropriately equipped online workbench of tools—to support field work on the documentation of endangered languages, to do research on translation, etc.

The front page of the infrastructure is displayed in Figure 1.

2.4. All users welcome

PORTULAN CLARIN favors and promotes Open Science, Open Access, Open Data and Open Source policies. Accordingly, all users are welcome to use and benefit from the scientific resources it distributes, with no user registration needed.

To ensure the quality of the scientific resources it distributes, depositors of resources are requested to register

before depositing and distributing their resources through the infrastructure. This is a very lean procedure, asking only for a user name, email and affiliation.

3. Repository

A major pillar in PORTULAN CLARIN mission is the distribution and preservation of language resources, including language data and language processing tools.

3.1. Deposit

Resource archival is ensured by maintaining a repository to which these scientific resources, together with their corresponding metadata information, may be deposited by registered users for long-term archiving and access, and from which any visitor can obtain copies of resources that are relevant for them.

Basic curation of the resources submitted to the repository is performed by checking the completeness and well-formedness of the metadata. The resource submission process relies on online forms, and on a workflow that ensures that the depositor is prompted when required information is lacking and that the required steps are performed for a submission to be completed and accepted.

After the metadata is submitted to the repository, the basic curation process is continued by the repository staff by means of manual assessment of the metadata and by means of checking its correspondence to the resource to be deposited.

Resource depositors are prompted, as it is in their best interest, to provide in the relevant metadata field a canonical citation for the resource being deposited.

Every resource in the repository is assigned a persistent identifier (PID) for long-term referencing.

Resources are being uploaded at a good pace by users, with the repository containing a large number (hundreds) of resources and growing.

3.2. Retrieval

The scientific resources stored and distributed through the repository can be searched by keyword match on the resource name and on its description, with faceted search bringing further filtering on metadata fields, such as the language, modality type, media type, etc.

Periodically, the metadata records are automatically harvested to the Virtual Language Observatory (VLO),[4] which acts as a central search hub for the whole, pan-European CLARIN ecosystem of repositories.

Note that the keyword search runs over the name and description metadata fields, not over the data content of the resource. Search over the data content of some resources is possible through the Federated Content Search (FCS) functionality of CLARIN.[5] This functionality allows running a query from a central location over multiple data sets, distributed over different national CLARIN nodes.

Figure 2(a) shows the search page of the PORTULAN CLARIN repository, with the list of resources ordered alphabetically. The text box on the top is used for keyword search, while the options of the right allow performing faceted search, which filters the results by multiple criteria (e.g. the language, the modality, etc). Figure 2(b) shows an example of a landing page for a resource.

3.3. Technological underpinnings

The repository is built with the Django[6] Web framework. The underlying database schema and workflow logic have been developed as an enhancement of the previously available METASHARE[7] repository software.

The repository website is created with the Bootstrap[8] CSS front-end framework, which provides a consistent and responsive interface that gracefully handles access from desktop and mobile platforms.

The keyword search functionality relies on Apache Solr[9] for efficient indexing.

The automatic metadata harvesting to the VLO central search hub is done using the OAI-PMH[10] protocol for repository interoperability.

4. Workbench

Another important part of PORTULAN CLARIN mission is to provide access to Language Technology tools and applications. This is accomplished through a workbench that makes available a wide range of processing tools and applications, whose display is grouped by categories, e.g. POS tagging, named entity recognition, sentiment analysis, etc. There are now about a couple of dozen services available, and their number is growing. A screenshot of the main workbench page is shown in Figure 3(a).

4.1. Modes of usage

The tools and applications are made available through different modes of interaction, namely through the browser, through file processing and through web services.

Using the tools and applications through the browser allows the user to directly enter the input, press a button and immediately get the result. Figure 3(b) shows an example of this mode of interaction.

While direct interaction through the browser is useful for short amounts of input, or as a demonstration of the capabilities and output format of a tool, large amounts of input need different modes of interaction. For relevant tools, the workbench also makes available a file processing mode of interaction which allows uploading files to be processed. The task of processing the uploaded files will be added to a queue and handled asynchronously. After the files are processed, the user will be notified by email and will be able to download the result from a unique URL generated when the task was submitted.

A third mode of interaction permits accessing the tools and applications as web services. This is particularly useful for end-users wanting to integrate some of the tools into their own processing workflow without having to be concerned with installing the tools locally on their own machines, or

[4] https://vlo.clarin.eu/

[5] https://contentsearch.clarin.eu/

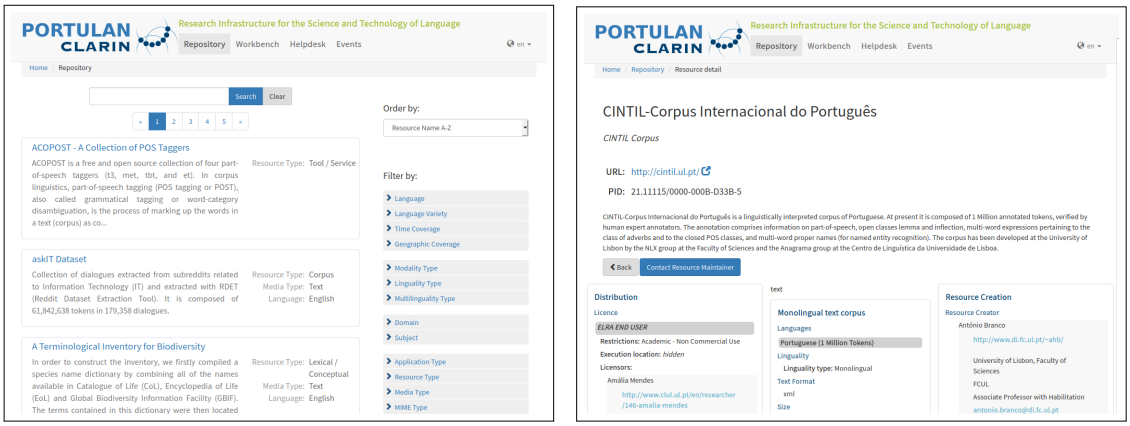
[6] https://www.djangoproject.com/

[7] http://www.meta-share.org/

[8] https://getbootstrap.com/

[9] https://lucene.apache.org/solr/

[10] https://www.openarchives.org/pmh/



(a) Search page of the repository

(b) Landing page of a resource

Figure 2: Screenshots of the repository

to depositors that wish to make the functionality of a tool accessible to end-users without releasing the tool itself.
Tools that are made available as a web service expose a programmatic interface that can be seamlessly invoked remotely. Their usage can be combined with the help of the CLARIN Language Resources Switchboard facility,[11] which provides a central location from where to find and connect webservices that are part of the wider CLARIN ecosystem.

4.2. Technological underpinnings

The tools and applications in the workbench may vary a lot in terms of the software used to implement them (C, Java, Perl, Python, etc.) and in terms of the supporting software libraries they require. To better cope with this heterogeneous environment, all tools and applications are organized into separate Docker containers. This greatly facilitates their configuration, minimizes system-wide dependencies and, by employing multiple instances of a container, allows performing load-balancing in a straightforward way.
Communication between tools is accomplished by the standard XML-RPC protocol. The same protocol is used for the web services.
The workbench website is also created with the Bootstrap framework, providing a consistent and responsive interface throughout the whole of PORTULAN CLARIN.

5. Help Desk and Consultancy

An important component of PORTULAN CLARIN mission is to provide support to the community of users of language technology. This is done through a help desk service for the infrastructure itself, and through a Language Technology consultancy service for the community at large.

5.1. Help desk

PORTULAN CLARIN staff runs a help desk that provides a user support service for the infrastructure, for the data sets in its repository and for the processing tools and services it makes available. This is useful for all users, but particularly suited for early career students and also for research from scientific domains with less ICT technical skills.
Besides providing help on how to use the scientific resources in the infrastructure and with troubleshooting issues, user support also involves the enhanced curation of submitted resources. This permits, for instance, to provide help in converting the deposited resources to formats other than their original formats, including standard formats, which should be particularly useful for users that lack the technical expertise to do the format conversion.

5.2. Consultancy

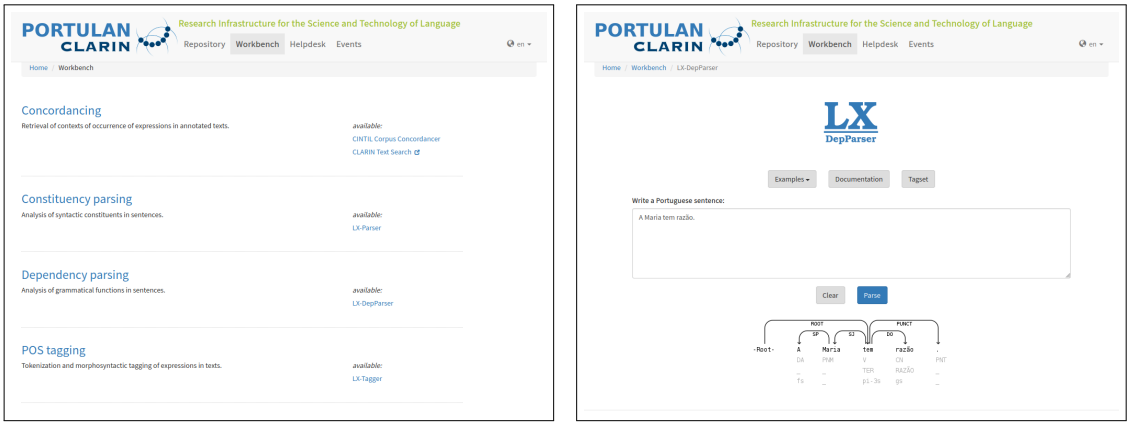
Another goal of CLARIN is to share knowledge, thus ensuring that the expertise that exists distributed over the various member countries of CLARIN is readily accessible, both within the infrastructure and to the research community as a whole. This is accomplished through the establishing of Knowledge Centres (K-centers), which are entities centrally certified by the CLARIN as being able to provide expert advice on some field.
PORTULAN CLARIN is recognized as a K-Centre specialized for the Science and Technology of the Portuguese Language, addressing all topics concerning this language: from Phonetics to Discourse and Dialogue; considering all language functions, from communicative performance to cultural expression; approached by all disciplines, from Theoretical Linguistics to Language Technology; covering all language variants, from national standard varieties across the world to dialects of professional groups; and taking into account all media of representation, from audio to brain imagiology recordings.

6. Depositing and Licensing

6.1. Deposit license

To deposit a resource, the user needs to fill in a respective metadata record and submit an instance of the deposit agreement template. This agreement grants a non-exclusive license for distribution of that resource to the PORTULAN

[11] https://switchboard.clarin.eu/



(a) Workbench page showing the list of tools

(b) A tool (dependency parser) used in-browser

Figure 3: Screenshots of the workbench

CLARIN research infrastructure, and therefore does not prevent the user from exercising their rights to distribute or publish the resource elsewhere.
This license is for distribution only, and therefore does not transfer the property or moral rights to the infrastructure.

6.2. Usage licenses

While PORTULAN CLARIN adheres to Open Science, Open Access, Open Data and Open Source policies, it does not impose them on its users.
In order to ensure the distribution of and access to the widest possible collection of scientific resources, the scientific resources in the repository are licensed by the respective depositors with the license set of their choice. This includes licensing resources for restricted usages, e.g. research, non commercial only, etc., and thus requiring that the end user proceeds to identify himself under the terms that may be required by the depositor.
When the depositor needs help in finding a suitable license for a resource, PORTULAN CLARIN provides support via its help desk and with online advice services like the CLARIN License Category Calculator.[12]
The license of a resource is stored as part of its metadata and is presented to any user attempting to have access to it. To eventually get access to a resource, a user has to explicitly accept the respective license. In order to obtain a copy of a resource with special restrictions or sensitive data, the user may be directed to the respective depositor in order to arrange for the compliance with the specific terms of that licensing.
The PORTULAN CLARIN repository provides long-term storage and distribution of data. The responsibility of following disciplinary and ethical norms for data storage and distribution lies with the repository. The responsibility of following disciplinary and ethical norms for the creation and gathering of data lies with the depositor of the data. As noted above, to deposit a resource in PORTULAN CLARIN, the depositor has to fill in and submit a depository agreement. In this agreement, it is explicitly stated that disciplinary and ethical norms were complied with when the resource was created. The depositor also has to specify whether the resource contains confidential data that could potentially be disclosed and the presence of such data will restrict the set of possible licenses that can be associated to the resource and end users that can have access to it.

7. Certification

PORTULAN CLARIN complies with the highest standards for research infrastructures. This is certified at different levels, by different entities.

7.1. International

PORTULAN CLARIN holds the international CoreTrustSeal[13] certification.[14] This certificates the compliance with a systematic range of organizational and technical requirements, such as its long-term sustainability plan, compliance to disciplinary and ethical norms, guarantees of data integrity and authenticity, software and hardware stability, data security, among many others.

7.2. European

As one of its national nodes, PORTULAN CLARIN is part of the CLARIN ERIC, which holds the European ESFRI-European Strategy Forum on Research Infrastructures[15] certification[16] as a landmark research infrastructure.[17]
Additionally, PORTULAN CLARIN holds the European

[12] https://www.clarin.eu/content/clarin-license-category-calculator

[13] https://www.coretrustseal.org/

[14] https://www.coretrustseal.org/wp-content/uploads/2019/12/PORTULAN-CLARIN.pdf

[15] https://www.esfri.eu/

[16] http://roadmap2018.esfri.eu/projects-and-landmarks/browse-the-catalogue/clarin-eric

[17] http://roadmap2018.esfri.eu/

CLARIN ERIC certification as a Knowledge Centre[18] and the European CLARIN ERIC certification as a national centre.[19]

7.3. National

FCT—Foundation for Science and Technology,[20] from the Portuguese Ministry of Science, Technology and Higher Education, is the national funding agency for scientific research. PORTULAN CLARIN holds the national certification from FCT as a research infrastructure of the National Roadmap of Research Infrastructures of Strategic Relevance.[21]

8. Governance and network

8.1. Network of implementation partners

The implementation of the infrastructure was undertaken under a project whose three core proponents partners are the Faculty of Sciences of the University of Lisbon, the School of Arts and Humanities of the University of Lisbon and the University of Évora.

Additionally, the implementation project is supported by a wide network of implementation partners. This network is open to further partners and currently encompasses over twenty research centers and organizations from the large range of scientific domains served by the infrastructure. There are partners from Brazil and Portugal, from all regions of Portugal, including the Azores islands. The Camões Institute, the Portuguese national organization responsible for the Portuguese language policy, is also part of the network and helps to pursue that part of the mission of the infrastructure concerned with the promotion of the Portuguese language.

The implementation partners are actively involved in depositing scientific resources and in the enhancing of the infrastructure. The list of implementation partners is open to further contributions and, as the infrastructure will evolve, it will include more organizations from all domains, including from the Humanities, Artificial Intelligence, Neuroscience, etc.

A list of the current network centers is provided in the Annex A.

8.2. Governance and staff

The infrastructure staff members have a large experience in the development of linguistic resources, data curation, natural language data processing, technical maintenance and software development. Most of them are also experts in the field of Language Technology who publish on, and attend, top-ranked scientific conferences in their domains of expertise.

The governance of the infrastructure includes a Board of Directors and a Management Team:

- Board of Directors
  - Director General: António Branco
  - Executive Director: Amália Mendes
  - Executive Director: Paulo Quaresma
- Management Team
  - Technical Manager: Luís Gomes
  - Scientific Resources and Users Support Manager: João Ricardo Silva
  - Communication and Administrative Manager: Andrea Teixeira

9. Conclusion

This paper presented the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language, which is the Portuguese national node of the pan-European research infrastructure CLARIN ERIC, with 20 member countries, and is part of the Portuguese national Roadmap of Research Infrastructures of Strategic Relevance.

Its mission is to support the widest range of users who need to resort to research results from the Science and Technology of Language. This is pursued through three main pillars in the infrastructure: (i) a *repository* for long-term archiving and access of language resources, be them language data or tools; (ii) a Language Technology *workbench* that makes available a wide range of language processing tools and applications, through various modes of interaction; and (iii) *help desk and consultancy* services that provide support to its users.

The infrastructure adheres to the principles of Open Science and its services are open to all users with no need of user registration or other dispensable access barriers.

Acknowledgements

The results reported here were partially supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa2020, Alentejo2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

Annex A Network of implementation partners

- Cristina Martins and Margarita Correia, Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Faculdade de Letras Universidade de Coimbra, Portugal
- Pilar Barbosa and Cristina Flores, Centro de Estudos Humanísticos (CEHUM), Universidade do Minho, Portugal
- Augusto Silva, Centro de Estudos Filosóficos 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

[18] https://www.clarin.eu/content/knowledge-centres

[19] https://www.clarin.eu/content/clarin-centres

[20] https://www.fct.pt/index.phtml.en

[21] https://www.fct.pt/apoios/equipamento/roteiro/index.phtml.en

- Amália Mendes, Centro de Linguística da Universidade de Lisboa (CLUL), Faculdade de Letras, Universidade de Lisboa, Portugal
- Fátima Oliveira, João Veloso and Rui Silva, Centro de Linguística da Universidade do Porto (CLUP), Faculdade de Letras, Universidade do Porto, Portugal
- Maria do Céu Caetano and Francisca Xavier, 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 (FCUL), Universidade de Lisboa, Portugal
- Paulo Quaresma, Laboratório de Ciência da Computação e Informática (NOVA LINCS), Instituto de Engenharia de Sistemas e Computadores (INESC), Escola de Ciências e Tecnologia, Universidade de Évora, Portugal
- Nuno Mamede, Instituto de Engenharia de Sistemas e Computadores (INESC), Instituto Superior Técnico, 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 and Nuno Marques, Laboratório de Ciência da Computação e Informática (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
- Eugénio Oliveira and Henrique Lopes Cardoso, Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal
- Vera Strube de Lima and Renata Vieira, Faculdade de Informática (FACIN), Pontifícia Universidade Católica do Rio Grande do Sul, Brasil
- Aline Villavicencio and Vera Strube de Lima, 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 — Instituto da Cooperação e da Língua, Portugal