

Explorar os dados com serviços de processamento respetivos: a experiência da infraestrutura PORTULAN CLARIN

António Branco, Amália Mendes e Paulo Quaresma



Repositório

Recursos científicos: dados e software

The screenshot shows a web browser window with the address bar displaying `https://portulanclarin.net/repository/search/`. The page header features the PORTULAN CLARIN logo and the tagline "Research Infrastructure for the Science and Technology of Language". Navigation links include "Repository", "Workbench", "Helpdesk", and "Outreach". A language selector is set to "en".

The main content area displays search results for the "Repository" section. A search bar with "Search" and "Clear" buttons is at the top. Below it, a pagination control shows page 1 of 5. The results are listed in three cards:

- U-Compare Sentence Splitting Service**
Web service created by exporting UIMA-based workflow from the U-Compare text mining system. Functionality: Identifies sentences in plain text Tools in workflow: Freeing sentence splitter web service (service provided by the PANACEA project) NOTE: The licence provided covers the web service o...
- The CIEMPIESS Proper-Names Pronouncing Dictionary**
Transcriptions in the CIEMPIESS-PNPD are based on a phonetic alphabet called Mexbet. Mexbet was design for the Spanish of Central Mexico and it has several levels of granularity. The CIEMPIESS-PNPD comes in two versions: Mexbet T29 and Mexbet T66. Level T29 of Mexbet means that transcriptions ...
- PS corpus (Post-Scriptum)-ES**
PS Corpus (Post-Scriptum)-ES is a corpus of 2368 informal mail letters written in Spanish during the Modern Ages (from the XVth century to the beginning of the XIXth century). Each letter is available as a semi-palaeographic transcription, a modernized transcription, and with part-of-speech a...

Each result card includes metadata: Resource Type (Tool / Service, Corpus), Media Type (Text), and Language (Welsh, Spanish; Castilian).

On the right side, there are two filter sections:

- Order by:** A dropdown menu currently set to "Language Name Z-A".
- Filter by:** A list of filter categories with expandable options:
 - Language
 - Language Variety
 - Time Coverage
 - Geographic Coverage
 - Modality Type
 - Linguality Type
 - Multilinguality Type
 - Domain
 - Subject
 - Application Type
 - Resource Type
 - Media Type
 - MIME Type

Bancada

Serviços online de processamento de dados

The screenshot shows the PORTULAN CLARIN website interface. The browser address bar displays the URL <https://portulanclarin.net/repository/search/>. The website header includes the PORTULAN CLARIN logo, the tagline "Research Infrastructure for the Science and Technology of Language", and navigation links for Repository, Workbench (highlighted with an orange circle), Helpdesk, and Outreach. A language dropdown menu is set to "en".

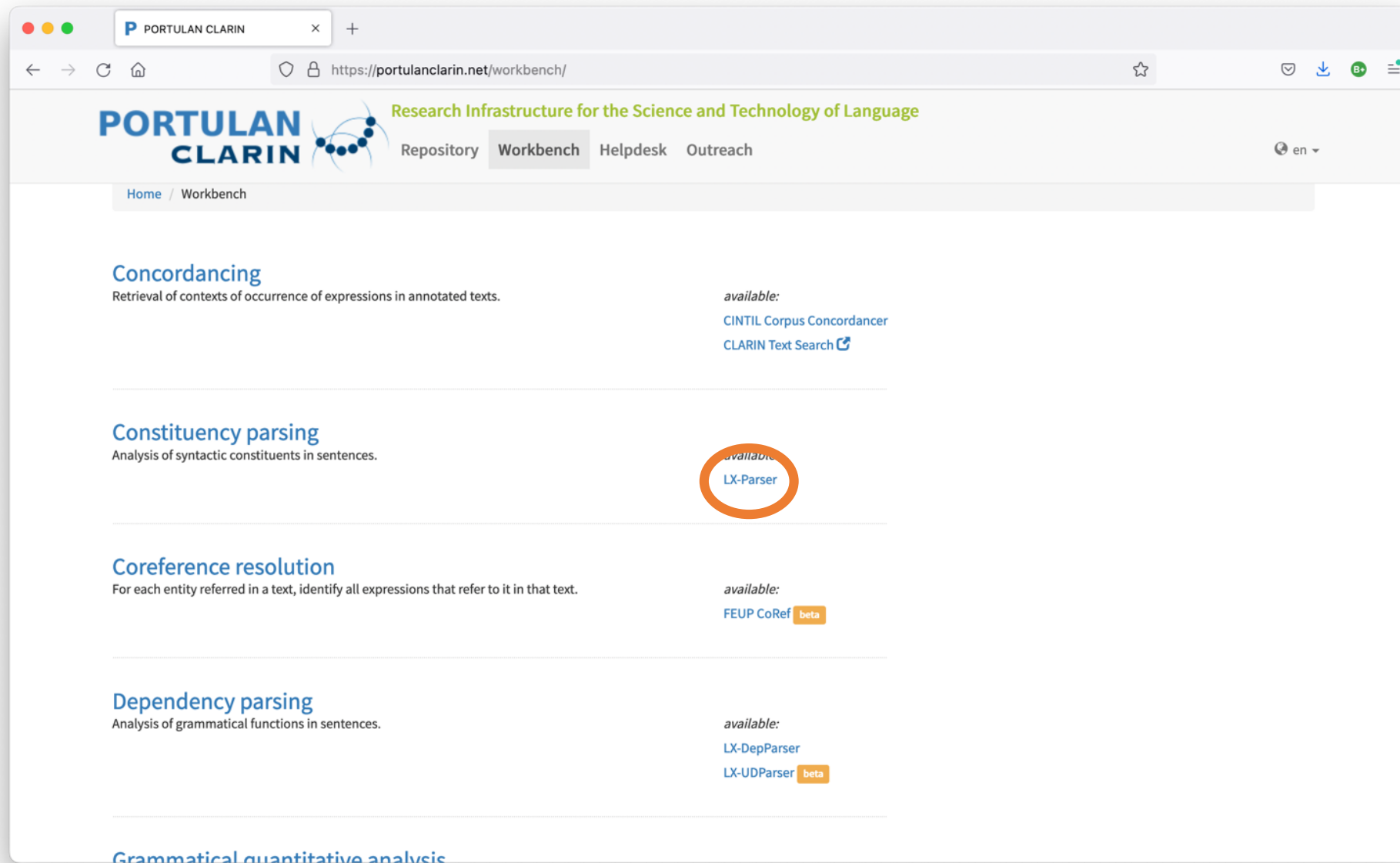
Below the header, a search bar with "Search" and "Clear" buttons is present. A pagination control shows page 1 of 5. The main content area displays three search results:

- U-Compare Sentence Splitting Service**
Web service created by exporting UIMA-based workflow from the U-Compare text mining system. Functionality: Identifies sentences in plain text Tools in workflow: Freeling sentence splitter web service (service provided by the PANACEA project) NOTE: The licence provided covers the web service o...
Resource Type: Tool / Service
Language: Welsh
- The CIEMPIESS Proper-Names Pronouncing Dictionary**
Transcriptions in the CIEMPIESS-PNPD are based on a phonetic alphabet called Mexbet. Mexbet was design for the Spanish of Central Mexico and it has several levels of granularity. The CIEMPIESS-PNPD comes in two versions: Mexbet T29 and Mexbet T66. Level T29 of Mexbet means that transcriptions ...
Resource Type: Corpus
Media Type: Text
Language: Spanish; Castilian
- PS corpus (Post-Scriptum)-ES**
PS Corpus (Post-Scriptum)-ES is a corpus of 2368 informal mail letters written in Spanish during the Modern Ages (from the XVIth century to the beginning of the XIXth century). Each letter is available as a semi-palaeographic transcription, a modernized transcription, and with part-of-speech a...
Resource Type: Corpus
Media Type: Text
Language: Spanish; Castilian

On the right side, there are filters for "Order by:" (set to "Language Name Z-A") and "Filter by:" with a list of expandable categories: Language, Language Variety, Time Coverage, Geographic Coverage, Modality Type, Linguality Type, Multilinguality Type, Domain, Subject, Application Type, Resource Type, Media Type, and MIME Type.

Tarefas de processamento

Execução online de cada tarefa na bancada



Serviço de processamento e suas interfaces online

Toda a gama de utilizadores: leigos a peritos

PORTULAN CLARIN / Workbench X

https://portulanclarin.net/workbench/lx-parser/

PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language

Repository Workbench Helpdesk Outreach

en

Home / Workbench / LX-Parser

LX Parser

Examples File Processing Notebook Web Service Documentation Tagset

Write a Portuguese sentence

Visualization format: ☒ friendly ☐ parentheses ☐ table

Clear Parse

Interface online: widget 1/5

PORTULAN CLARIN / Workbench X

https://portulanclarin.net/workbench/lx-depparser/

PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language

LX DepParser

Examples File Processing Notebook Web Service Documentation Tagset

A emissão será cotada na Bolsa de Valores do Luxemburgo.

Visualization format: ☒ friendly ☐ CONLL Tagset: ☒ CINTIL ☐ universal dependencies

Clear Parse

Diagram illustrating the dependency parse for the sentence "A emissão será cotada na Bolsa de Valores do Luxemburgo."

The diagram shows the following structure:

- ROOT (A) is the root of the tree.
- SP (emissão) is the subject.
- SJ (será) is the main verb.
- PRD (cotada) is the object.
- OBL (em_) is the object of the preposition "na".
- C (a) is the object of the preposition "de".
- SP (Bolsa) is the object of the preposition "de".
- N (de) is the object of the preposition "de".
- N (Valores) is the object of the preposition "de".
- PREP (de_) is the object of the preposition "de".
- DA (o) is the object of the preposition "de".
- PNM (Luxemburgo) is the object of the preposition "de".
- PUNCT (.) is the end of the sentence.

The diagram also shows the following dependencies:

- ROOT (A) is connected to SP (emissão) via SP.
- ROOT (A) is connected to SJ (será) via SJ.
- ROOT (A) is connected to PRD (cotada) via PRD.
- ROOT (A) is connected to OBL (em_) via OBL.
- ROOT (A) is connected to C (a) via C.
- ROOT (A) is connected to SP (Bolsa) via SP.
- ROOT (A) is connected to N (de) via N.
- ROOT (A) is connected to N (Valores) via N.
- ROOT (A) is connected to PREP (de_) via PREP.
- ROOT (A) is connected to DA (o) via DA.
- ROOT (A) is connected to PNM (Luxemburgo) via PNM.
- ROOT (A) is connected to PUNCT (.) via PUNCT.

Anotação morfossintática

Análise de sentimentos

Análise gramatical quantitativa

Análise sintática de constituição

Análise sintática de dependências

Análise sub-sintática

Análise temporal

Busca em treebank

Classificação de proficiência

Concordanceamento

Busca de Texto na CLARIN

Consulta da wordnet

Estilometria

Etiquetador de papéis semânticos

Extração de palavras chave

Flexão nominal

Lematização verbal

Lematização e flexão de expressões verbais.

Normalização ortográfica

Reconhecimento de entidades nomeadas

Resolução de correferência

Semelhança semântica

Separação de frases

Separação de unidades lexicais

Silabificação

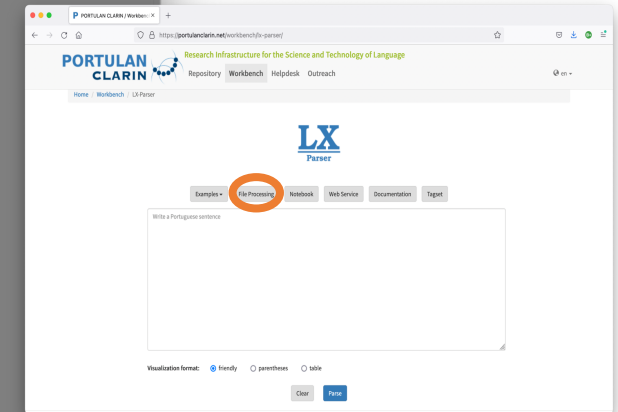
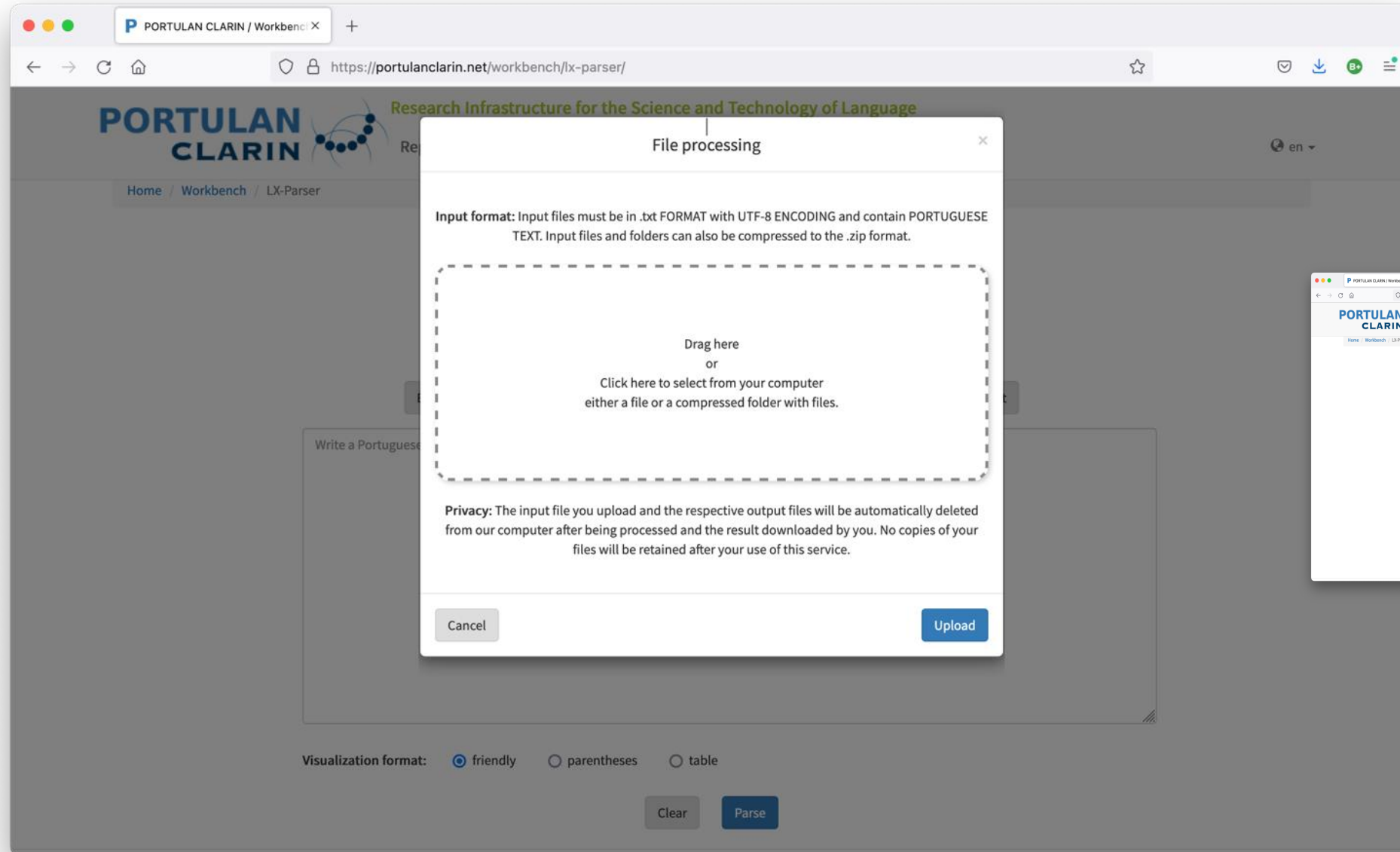
Silabificação de expressões.

Tradução

Transcrição

Transcrição fonológica

Interface online: file processing 2/5



Interface online: jupyter notebook 3/5

Using LX-Parser to parse sentences and displaying constituency trees

This is an example notebook that illustrates how you can use the LX-Parser web service to parse sentences.

Before you run this example, replace `access_key_goes_here` by your webservice access key, below:

In [1]:

```
LXPARSER_WS_API_KEY = 'access_key_goes_here'
LXPARSER_WS_API_URL = 'https://portulanclarin.net/workbench/lx-parser/api/'
```

Importing required Python modules

The next cell will take care of installing the `requests`, `nlk` and `svglng` packages, if not already installed, and make them available to use in this notebook.

In [2]:

```
try:
    import requests
except:
    !pip3 install requests
    import requests
try:
    import nltk.tree
except:
    !pip3 install nltk
    import nltk.tree
try:
    import svglng
except:
    !pip3 install svglng
    import svglng

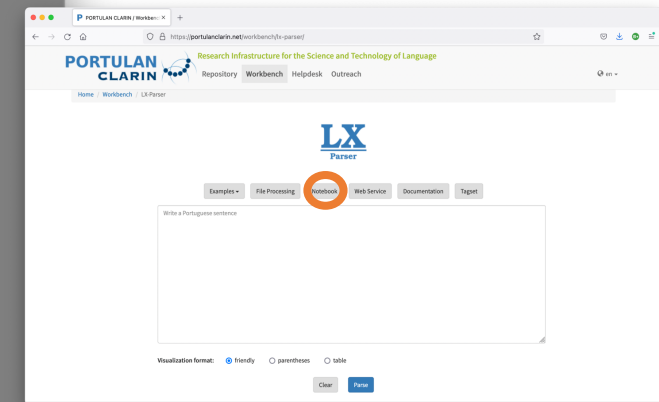
import IPython.display
```

Wrapping the complexities of the JSON-RPC API in a simple, easy to use function

The `WSEException` class defined below, will be used later to identify errors from the webservice.

In [3]:

```
class WSEException(Exception):
    'Webservice Exception'
    def __init__(self, errordata):
```



Interface online: webservice 4/5

The screenshot displays the PORTULAN CLARIN LX-Parser web service interface. The main window shows instructions on how to use the web service, including a code example for using the service with Python 3 and the requests library. A modal window titled "Instructions to use this web service" is open, providing detailed information about the service's availability, usage, and a code example. The code example shows how to install the requests library and how to use the LX-Parser web service to parse a Portuguese sentence. The interface also includes a "Request key" button and a "Web Service" tab in the navigation menu.

PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language

Home / Workbench / LX-Parser

Instructions to use this web service

The web service for this application is available at <https://portulanclarin.net/workbench/lx-parser/api/>.

Below you find an example of how to use this web service with Python 3.

This example resorts to the requests package. To install this package, run this command in the command line: `pip3 install requests`.

To use this web service, you need an access key you can obtain by clicking in the button below. A key is valid for 31 days. It allows to submit a total of 10 million characters by means of requests with no more 2500 characters each. It allows to enter 100,000 requests, at a rate of no more than 200 requests per hour.

[Request key](#)

For other usage regimes, you should contact the [helpdesk](#).

The input data and the respective output will be automatically deleted from our computer after being processed. No copies will be retained after your use of this service.

```
import json
import requests # to install this library, enter in your command line:
                 # pip3 install requests

# This is a simple example to illustrate how you can use the LX-Parser web service

# Requires: key is a string with your access key
# Requires: text is a string, UTF-8, with a maximum 2500 characters, Portuguese text, with
#           the input to be processed
# Requires: format is a string, indicating the output format, which can be either
#           'parentheses', 'table' or 'JSON'

# Ensures: output according to specification in https://portulanclarin.net/workbench/lx-parser/
# Ensures: dict with number of requests and characters input so far with the access key, and
#           its date of expiry

key = 'access_key_goes_here' # before you run this example, replace access_key_goes_here by
                             # your access key

format = 'parentheses' # other possible values are 'table' and 'JSON'
```

PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language

Repository Workbench Helpdesk Outreach

Home / Workbench / LX-Parser

LX Parser

Examples File Processing Notebook **Web Service** Documentation Taget

Write a Portuguese sentence

Visualization format: ☒ friendly ☐ parentheses ☐ table

[Clear](#) [Parse](#)

Repositório: código 5/5

The screenshot shows a web browser window with the URL <https://portulanclarin.net/repository/search/>. The page header features the PORTULAN CLARIN logo and the tagline "Research Infrastructure for the Science and Technology of Language". Navigation links include Repository, Workbench, Helpdesk, and Outreach. A language selector is set to "en".

The main content area displays search results for the repository. A search bar with "Search" and "Clear" buttons is at the top. Below it, a pagination bar shows results 1 through 5, with "1" being the current page. The results are listed in three cards:

- U-Compare Sentence Splitting Service**
Web service created by exporting UIMA-based workflow from the U-Compare text mining system. Functionality: Identifies sentences in plain text Tools in workflow: Freeing sentence splitter web service (service provided by the PANACEA project) NOTE: The licence provided covers the web service o...
- The CIEMPIESS Proper-Names Pronouncing Dictionary**
Transcriptions in the CIEMPIESS-PNPD are based on a phonetic alphabet called Mexbet. Mexbet was design for the Spanish of Central Mexico and it has several levels of granularity. The CIEMPIESS-PNPD comes in two versions: Mexbet T29 and Mexbet T66. Level T29 of Mexbet means that transcriptions ...
- PS corpus (Post-Scriptum)-ES**
PS Corpus (Post-Scriptum)-ES is a corpus of 2368 informal mail letters written in Spanish during the Modern Ages (from the XVth century to the beginning of the XIXth century). Each letter is available as a semi-palaeographic transcription, a modernized transcription, and with part-of-speech a...

Each result card includes metadata: Resource Type (Tool / Service, Corpus), Media Type (Text), and Language (Welsh, Spanish; Castilian).

On the right side, there are two filter sections:

- Order by:** A dropdown menu currently set to "Language Name Z-A".
- Filter by:** A list of filter categories with expandable options:
 - Language
 - Language Variety
 - Time Coverage
 - Geographic Coverage
 - Modality Type
 - Linguality Type
 - Multilinguality Type
 - Domain
 - Subject
 - Application Type
 - Resource Type
 - Media Type
 - MIME Type

Repositório x Bancada

Explorar e enriquecer os dados com serviços de processamento respetivos

The image shows a screenshot of the PORTULAN CLARIN website. The browser's address bar displays "https://portulanclarin.net". The website's header includes the PORTULAN CLARIN logo and the tagline "Research Infrastructure for the Science and Technology of Language". A navigation menu contains the following items: "Repository", "Workbench", "Helpdesk", and "Outreach". The "Repository" and "Workbench" items are circled in orange. The main content area features a dark blue background with various scientific and linguistic visualizations, including brain scans, linguistic trees, and mathematical formulas. A large, bold, white text overlay reads "DISTRIBUTE REUSE FOSTER". Below this, a smaller text overlay states: "Services and data for researchers, innovators, students and language professionals." The website also includes a language selector set to "en".

PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language

Repository Workbench Helpdesk Outreach

DISTRIBUTE REUSE FOSTER

Services and data for researchers, innovators, students and language professionals.

obrigado
portulanclarin.net

António Branco, Amália Mendes e Paulo Quaresma

