

Wikisource em língua portuguesa: edição colaborativa de recursos lexicográficos

Ana Salgado, Leonor Martins, Leonor Reis,
Carlos Silva, Luís Trigo, André Barbosa

Festa do Software Livre
Faculdade de Engenharia da Universidade do Porto
4 de outubro



U. PORTO
FLUP FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

 Centro de
Linguística da
Universidade do
Porto

 **CODA**
Centre for
Digital Culture
and Innovation

fct Fundação
para a Ciência
e a Tecnologia


WIKIMEDIA
PORTUGAL

Agenda

- Contexto
 - ◆ Humanidades Digitais, Lexicografia e acesso aberto
- O que é a Wikisource?
 - ◆ Biblioteca digital colaborativa
- Estudo de caso (Dalgado, 1913)
 - ◆ *Influência do Vocabulário Português nas Línguas Asiáticas*
- Fluxo de trabalho
 - ◆ Do OCR à transclusão
- Métricas
 - ◆ Avaliação quantitativa
- Desafios
 - ◆ Questões em aberto
- Considerações finais

Objetivo

Explorar como a **Wikisource** pode facilitar a **preservação e a democratização de recursos lexicográficos históricos**, através de metodologias colaborativas e tecnologias de software livre.

Contexto



Humanidades digitais

As Humanidades Digitais permitem ultrapassar as limitações das edições impressas e abrir novas formas de acesso.

As plataformas colaborativas tornam possível superar essas limitações, promovendo a democratização do conhecimento.



Acesso aberto

O movimento de acesso aberto assegura que o património textual em língua portuguesa esteja livremente disponível a investigadores, estudantes e ao público em geral.

Licenças como as Creative Commons garantem a sua reutilização aberta e sustentável.



Colaboração e ciência cidadã

A participação de voluntários e investigadores dá origem a uma comunidade dinâmica, comprometida com a preservação cultural.

A ciência cidadã integra a sociedade civil no processo científico, reforçando os laços entre a academia e o público > **ambiente wiki**.

O que é a Wikisource?

→ Biblioteca digital livre mantida pela Fundação Wikimedia

https://pt.wikisource.org/wiki/Wikisource:Página_principal

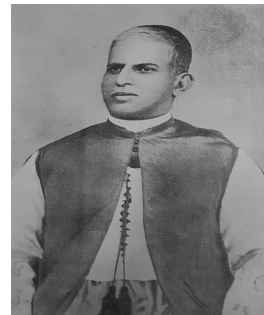
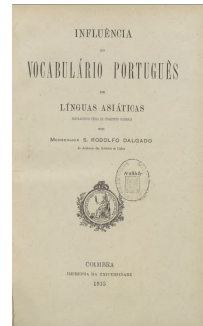
→ Dedicada à preservação e disponibilização de textos em domínio público ou sob licenças abertas compatíveis.

→ Funciona como um repositório colaborativo, onde voluntários de todo o mundo transcrevem, reveem e publicam obras literárias, científicas e históricas.

Estudo de caso

Influência do Vocabulário Português em Línguas Asiáticas

Sebastião Rodolfo Dalgado, 1913



Sebastião Rodolfo Dalgado (Goa, Goa Norte, Bardez, Assagão, 8 de Maio de 1855 — Lisboa, 4 de Abril de 1922)

- A obra analisa a integração de vocábulos portugueses em cerca de cinquenta idiomas asiáticos, evidenciando o impacto linguístico do período de expansão portuguesa na Ásia.
- **Porque não carregar a esta obra na Wikisource?**
 - ◆ O objetivo era tornar o recurso lexicográfico acessível a investigadores de todo o mundo, permitindo estudos linguísticos e comparativos que antes seriam impraticáveis.

Fluxo de trabalho



Criação da galeria-índice e metadados

Estruturação inicial do projeto com informações bibliográficas completas



OCR (Reconhecimento ótico de caracteres)

Processamento automático do ficheiro digitalizado para extração de texto



Normas de transcrição e formatação

Aplicação de regras editoriais para garantir consistência e fidelidade



Revisão colaborativa (ProofreadPage)

Validação por múltiplos colaboradores utilizando sistema de cores



Ligações à Wikidata e à Wikidata Lexemes

Conexão com dados estruturados para enriquecimento semântico



Transclusão e finalização

Publicação no espaço principal com navegação e categorização

Galeria-índice

Estruturação do projeto e preparação inicial

1

Verificar existência prévia

Antes de iniciar um projeto, é fundamental confirmar que o texto ainda não foi transcrito na Wikisource, evitando duplicação e garantindo eficiência.

2

Confirmar estado legal

Verificação cuidada de que a obra está em domínio público ou disponível sob licença compatível (Creative Commons ou similar), assegurando conformidade legal para disponibilização pública.

3

Carregamento da digitalização

Carregamento do ficheiro digitalizado para o Wikimedia Commons, repositório centralizado que permite o acesso e reutilização das digitalizações por todos os projetos Wikimedia.

4

Criação de metadados

Preenchimento da página Galeria com informações bibliográficas completas: título, autor(es), ano de publicação, editor, número de páginas e outras informações relevantes para catalogação.

5

Marcação <pagelist>

Configuração da numeração de páginas, permitindo distinguir entre numeração física (do documento digitalizado) e numeração lógica (conforme numeração original da obra impressa).

Galeria-índice



Search Wikimedia Commons

English Anacastroalgaldo



Wikimania 2026 scholarships are now open!
This application is open until Friday, 31st October 2025

Main menu

Main page

Welcome

Community portal

Village pump

Help center

Switch to old look

Participate

Upload file

Recent changes

Latest files

Random file

Contact us

Special pages

Contents

Beginning

Summary

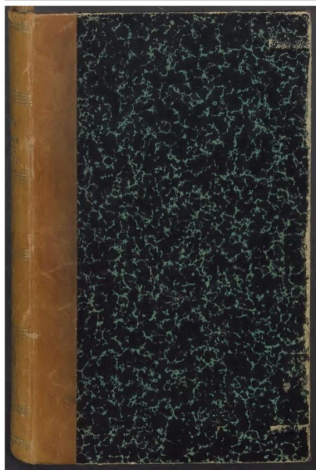
Licensing

File:Sebastião Rodolfo Dalgado - Influência do Vocabulário Português em Línguas Asiáticas (1913).pdf

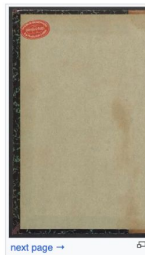
File Discussion

Read Edit View history

Download Use this file Use this file Email a link Information



Go to page 1



Size of this JPG preview of this PDF file: 405 × 599 pixels. Other resolutions: 162 × 240 pixels 1 325 × 480 pixels 1 1 006 × 1 487 pixels.

Original file (1,006 × 1,487 pixels, file size: 17.99 MB, MIME type: application/pdf, 355 pages)

Tools

Actions

Move

General

What links here

Related changes

Permanent link

Page information

Cite this page

Get shortened URL

Download QR code

Concept URI

Nominate for deletion

Print/export

Download as PDF

Printable version

Appearance

Text

Small

Standard

Large

This page always uses small font size

Width

Título

Influência do Vocabulário Português em Línguas Asiáticas

Subtítulo

Abrangendo cerca de cinquenta idiomas

Autor

Sebastião Rodolfo Dalgado

Editor

Imprensa da Universidade de Coimbra

Local

Coimbra

Editora

Imprensa da Universidade

Ano

1913

OCR + HTR

Reconhecimento automático de texto

A Wikisource integra ferramentas de OCR (*Optical Character Recognition*) para texto impresso, como o Tesseract e Google OCR, e HTR (*Handwritten Text Recognition*) para texto manuscrito, com o *Transkribus**, permitindo a extração automática de texto a partir das imagens digitalizadas.

* Os utilizadores das plataformas Wikimedia Lusófona têm acesso a créditos grátis *Transkribus*

Boas práticas colaborativas

Para otimizar o trabalho em equipa, recomenda-se dividir o texto por capítulos ou secções, permitindo que múltiplos voluntários trabalhem simultaneamente sem conflitos.

É fundamental que cada colaborador respeite as convenções editoriais estabelecidas, mantendo consistência ao longo de todo o documento.

Normas de transcrição

Garantindo fidelidade e consistência editorial

Predefinições usuais

A Wikisource disponibiliza predefinições (*templates*) para elementos comuns: texto centrado, disposição em colunas, notas de rodapé, texto em versaletes. Estas ferramentas garantem formatação consistente e facilitam a reprodução da estrutura original.

Gestão de translineação

Palavras divididas entre linhas devem ser reunidas mantendo a grafia correta. A numeração de páginas segue a marcação `<pagelist>` estabelecida, permitindo correspondência exata entre texto transcrito e original digitalizado.

Marcação de erros

Erros tipográficos ou grafias antigas presentes no original são preservados utilizando a predefinição `{{sic}}`, que indica ao leitor que a forma apresentada é fiel ao documento fonte, mesmo que não corresponda à norma atual.

Notas de rodapé

As notas devem ser transcritas mantendo a sua numeração e posição relativa. Quando uma nota se estende por múltiplas páginas, utilizam-se predefinições específicas para garantir continuidade e referência correta.

Critério de colunas

A decisão de manter ou desfazer a disposição em colunas baseia-se em critérios funcionais: se a apresentação em coluna única facilitar a leitura digital sem perder informação estrutural, pode optar-se por essa simplificação.

Revisão colaborativa

Sistema ProofreadPage

O sistema de revisão da Wikisource baseia-se numa metodologia de validação progressiva, utilizando códigos de cor para indicar o estado de cada página:

Cinzentos: página sem texto

Vermelho: página não revista

Amarelo: primeira revisão concluída

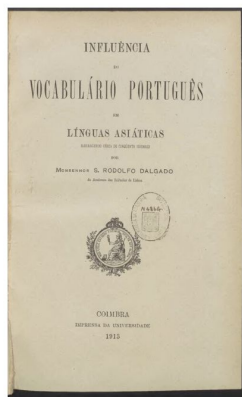
Verde: validado por segundo revisor

Azul: página com problemas identificados

Validação

- A metodologia de dupla revisão independente é fundamental para garantir a qualidade do texto transcrito.
- Um primeiro colaborador faz a revisão inicial, corrigindo erros de OCR e aplicando as normas de transcrição.
- Um segundo revisor valida o trabalho, identificando eventuais inconsistências ou erros remanescentes.

Este processo colaborativo assegura que o texto final apresente elevada fidelidade ao original, minimizando erros e preservando elementos estruturais e tipográficos relevantes.



Título Influência do Vocabulário
Português em Línguas Asiáticas

Subtítulo Abrangendo cerca de cinquenta idiomas

Autor Sebastião Rodolfo Dalgado

Editor Imprensa da Universidade de Coimbra

Local Coimbra

Editora Imprensa da Universidade

Ano 1913

Fonte Digitalização dos originais

Progresso Revisão pendente

Páginas	Capa	—	—	—	Frontispício	—
	Rosto	Licença	Dedicatória	—		
	VII	VIII	IX	Mapa linguístico	—	
	XI	XII	XIII	XIV	XV	XVI
	XVII	XVIII	XIX	XX	XXI	XXII
	XXIII	XXIV	XXV	XXVI	XXVII	XXVIII
	XXIX	XXX	XXXI	XXXII	XXXIII	
	XXXIV	XXXV	XXXVI	XXXVII		
	XXXVIII	XXXIX	XL	XL	XLII	
	XLIII	XLIV	XLV	XLVI	XLVII	
	XLVIII	XLIX	L	LI	LII	LIV
	LVI	LVII	LVIII	LIX	LX	LXI
	LXII	LXIII	LXIV	LXV	LXVI	LXVII
	LXVIII	LXIX	LXX	LXXI	LXXII	
	LXXIII	LXXIV	LXXV	LXXVI		
	LXXVII	LXXVIII	LXXIX	LXXX		
	LXXXI	LXXXII	LXXXIII	LXXXIV		
	LXXXV	LXXXVI	LXXXVII			
	LXXXVIII	LXXXIX	XC	XCI	XCII	

ÍNDICE

Parecer	VII
Prefácio	XI
Introdução:	
I. — Influência de Portugal no	
Oriente	XV
II. — Influência da língua	
portuguesa	XVIII
III. — Línguas da Índia, Noções	
gerais	XXI
IV. — Classificação e divisão	XXIII
V. — Distribuição geográfica	XXV
VI. — Objecto do estudo	XXVI
VII. — Elementos exóticos	XXVII
VIII. — Veículos e motivos da	
influência do português	XXVIII
IX. — Natureza morfológica dos	
exotismos	XXIX
IX. — Observações fonológicas .	XXX
X. — Observações fonológicas ..	XXX
XI. — Fontes e dificuldades do	
estudo	XXXII
XII. — Organização do	
Vocabulário	XXXIV
XIII. — Noções individuais das	
línguas	XXXVI
XIV. — Alfabetos e transcrições .	LXXI
Bibliografia	LXXIX
Relação alfabética das línguas e dos	
dialectos do Vocabulário	LXXXVI
Ordem das línguas na inscrição	
vocabular	LXXXIX
Abreviaturas	XCI
Vocabulário	I
Suplemento	165
Lista geral	167
Listas particulares	191
Erratas e aditamentos	251

Esta página foi validada

VOCABULÁRIO

A

Abada (port. ant.: «rinoceronte, rémea do rinoceronte»^[1]). Indo-
ingl. *abada* (obsol.).

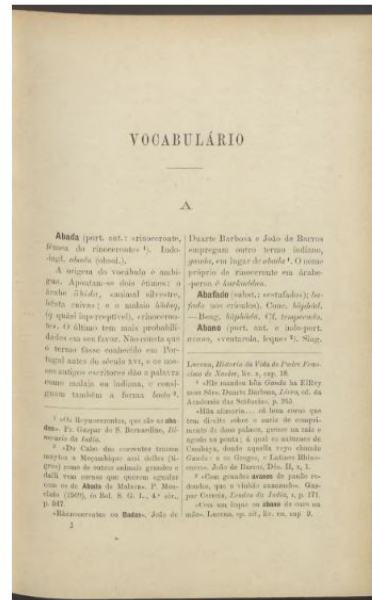
A origem do vocábulo é ambígua. Apontam-se dois étimos: o
arabe *ábida*, «animal silvestre, bête ruiva»; e o malaio *badaq*,
(q quase imperceptível), «rinoceronte». O último tem mais
probabilidades em seu favor. Não consta que o termo fôsse
conhecido em Portugal antes do século XVI, e os nossos antigos
escritores dão a palavra como malaia ou indiana, e consignam
também a forma *bada*^[2]. Duarte Barbosa e João de Barros
empregam outro termo indiano, *ganda*, era lugar de *abada*.^[3] O
nome próprio de rinoceronte em árabe-persa é *karkaddan*.

Abafado (subst.: «estufado»); *bafado* nos crioulos. Conc.
bâphâd. — Beng. *bâphâd*. Cf. *temperado*.

Abano (port. ant. e indo-port. *avano*, «ventarola, leque»^[4]).
Sing.

Notas

- ¹ ↑ «Os Reynocerontes, que são as *abadas*». Fr. Gaspar de S.
Bernardino, *Itinerário da Índia*.
- ² ↑ «Do Cabo das correntes trazem muytos a Moçambique assi
delles (figes) como de outros animais grandes e dalli vem
como que querem equalar com os de *Abada* de Malaca». P.
Mondalio (1569), in Bol. S. G. L., 4.ª sér., p. 547. «Rhenocerontes
ou *Badas*». João de Lucena, *História da Vida do Padre
Francisco de Xavier*, liv. x, cap. 18.
- ³ ↑ «Ele mandou hũa *Ganda* ha ElRey noso Sôr». Duarte Barbosa,
Livro, ed. da Academia das Ciências, p. 263. «Hũa almaria... cõ
hum como que tem dreito sobre o nariz de comprimento de
dous palmos, grosso na raiz e agudo na ponta; à qual os
naturaes de Cambaya, donde aquella veyo chamão *Ganda*; e os
Gregos, e Latinos Rhinoceros». João de Barros, Déc. II, x, f. 1.
- ⁴ ↑ «Com grandes *avanos* de pauão redondos, que o vinhão
auanando». Gaspar Correia, *Lendas da Índia*, i, p. 171. «Com
um leque ou *abano* de ouro na mão». Lucena, op. cit., liv. VII,
cap. 9.



Revisão do Reconhecimento de Texto

Ferramentas de OCR e HTR

No caso da obra de Dalgado, o Google OCR revelou melhor desempenho.

Correção manual

Para garantir fidelidade absoluta ao texto original, procedeu-se à correção manual de todas as incongruências na transcrição automática:

- caracteres de outros alfabetos incorretamente interpretados
- excertos em línguas estrangeiras
- espaçamentos irregulares
- quebras de linha inadequadas.

Preservação da formatação

Todos os elementos de formatação presentes no original são meticulosamente preservados:

- texto em itálico
- negritos,
- translineações
- estruturas especiais como notas de rodapé ou citações.

Novena. Conc. *novén'*. — Beng. *novená*. — Tam. *novenei*. — Tel. *novéna*. — Can. *novénu*.

Número. Conc. *num'r*, *numbr*, Term. vern. *ánk*, *sankhyá*, *gan*,

gantí. — ? Sing. *nómare*, *nomma-
raya* (talvez do ingl. *number*). —
? Bug. *nómore*; provavelmente do
hol. *nommer*. — Tet., Gal. *númeru*.
Term. vern. *súra*.

Antes do tratamento manual

114

OFENDER

OLA

Novena. Conc. *novérí*. — Beng. *novená*. — Tam. *novenei*. — Tel. *novéna*. — Can. *novénu*. Número. Conc. *num'r*, *numhr*, Term. vern. *ánk*, *sankhyá*, *gan*,

gantí. — ? Sing. *nómare*, *nommaraya* (talvez do ingl. *number*). — ? Bug. *nômore*; provavelmente do hol. *nommer*. — Tet., Gal. *númeru*. Term. vern. *súra*.

Após o tratamento manual

114

OFENDER

OLA

Novena. Conc. *novén*. — Beng. *novená*. — Tam. *novenei*. — Tel. *novéna*. — Can. *novénu*.

Número. Conc. *num'r*, *numbr*, Term. vern. *ánk*, *sankhyá*, *gan*, *gantí*. — ? Sing. *nómare*, *nomma-
raya* (talvez do ingl. *number*). — ? Bug. *nómore*; provavelmente do hol. *nommer*. — Tet., Gal. *númeru*. Term. vern. *súra*.

Transclusão e finalização



Passagem para o espaço principal

Após conclusão da revisão em todas as páginas, o texto é transcluído — transferido do espaço de trabalho ProofreadPage para o espaço principal da Wikisource, tornando-se acessível como obra completa e navegável.



Categorias e navegação

Atribuição de categorias temáticas e cronológicas que facilitam a descoberta do texto. Criação de ligações internas entre capítulos, sumários e índices, permitindo navegação intuitiva semelhante a um livro físico mas com as vantagens do hipertexto digital.



Revisão final de consistência

Verificação global da coerência entre capítulos, correção de referências cruzadas, confirmação de que todos os sumários estão completos e de que a numeração de páginas corresponde ao original. Esta revisão holística garante que a obra funciona como um todo integrado.

Wikidata Lexemes

Enriquecimento semântico através de dados estruturados

- Camada adicional de valor informativo, permitindo estabelecer ligações entre os lexemas identificados no texto e os dados estruturados disponíveis na base de dados universal da Wikimedia.
- Quando um lema já existe na Wikidata, cria-se uma ligação direta com o lexema correspondente utilizando um identificador único e permanente.
- Este sistema garante que a referência se mantém válida mesmo que o nome do item seja alterado ou traduzido.

`[[ :d:Lexeme:L1417720|Adeus ]]`

L1417720 — identificador único e permanente na Wikidata

Adeus — forma visualizada pelo utilizador no texto

Esta página ainda não foi revisada

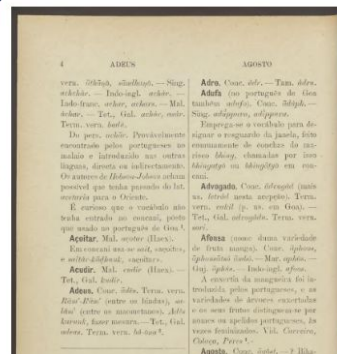
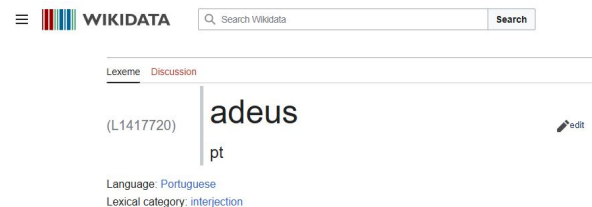
ADEUS

vern. *āthānō, sāndhanō*. - Sing. *achchār*. Indo-ingl. *achār*. Indo-franc. *achār, achars*. Mal. *āchar*. Tet., Gal. *achār, achār*. Term. vern. *budū*. Do pers. *achār*. Provavelmente encontrado pelos portugueses no malaio e introduzido nas outras línguas, directa ou indirectamente. Os autores de *Hobson-Jobson* acham possível que tenha passado do lat. *acetaria* para o Oriente. É curioso que o vocábulo não tenha entrado no concani, posto que usado no português de Goa^[1].

Açoitar. Mal. *açotar* (Haex). — Tet., Gal. *kudir*. Em concani usa-se *sait*, «açoitte» e *saitār-kādhunk*, «açoitars».

Acudir. Mal. *cudir* (Haex). — Tet., Gal. *kudir*.

Adeus. Conc. *ādēs*. Term. vern. *Rām-Rām* (entre os hindus), *sa-lām* (entre os maometanos). Adēs karunk, *fazer mesura*. - Tet., Gal. *adeus*. Term. vern. *bā-ōna*^[2].



Wikidata Lexemes

Conexão com dados estruturados para investigação avançada

```
'''[[d:Lexeme L477163 Número.]]''' Conc. ''num'  
Term. vern. ''ánk'', ''sankhyá'', ''gaṇ'',  
''gaṇtí''. — ? Sing. ''nómare'', ''nomma-''  
''raya'' (talvez do ingl. ''number''). —
```

Número. Conc. *num'r*, *numbr*, Term. vern. *ánk*, *sankh*,
— ? Sing. *nómare*, *nomma- raya* (talvez do ingl. *numt*
nómore; provavelmente do hol. *nommer*.—Tet., Gal. *n*
vern. *súra*.

Lexeme Discussion

(L477163)

número

pt

Ligação da obra

A obra completa é registada na Wikidata como entidade bibliográfica.

Associações lexicais

Estabelecem-se ligações entre lemas → obra → páginas específicas, criando uma rede semântica.

Potencial investigativo

Esta estruturação abre possibilidades para estudos lexicográficos computacionais: análises de frequência, estudos etimológicos comparativos, mapeamento de influências linguísticas e visualização de relações entre vocábulos.

Propriedades Wikidata

P304 (número de páginas) como qualificador de citações, **P7855** (attested as) para documentar variantes históricas de grafia, permitindo rastrear a evolução ortográfica de cada vocábulo.

Métricas

Avaliação quantitativa do trabalho realizado em 2 meses

211

Páginas transcritas

Total de páginas da obra de
Dalgado processadas e
disponibilizadas na plataforma
Wikisource

56

Páginas validadas

Páginas que passaram pelo
processo completo de dupla
revisão e foram marcadas
como finalizadas

3

Colaboradores ativos

Número de colaboradores que
contribuíram significativamente
para a transcrição e revisão do
texto

20 min

Tempo médio por página

Duração média necessária para
transcrever, formatar e validar
uma página completa, incluindo
ligações semânticas

Desafios

Predefinições específicas

Faltam modelos (*templates*) para tipologias textuais específicas: estruturas de dicionários, vocabulários, glossários.

Normalização editorial

Definir convenções sobre aspetos como: tratamento de variantes ortográficas históricas, critérios para simplificação de estruturas complexas em ambiente digital, e metodologias padronizadas para ligação de recursos lexicográficos à Wikidata.

Formação de colaboradores

Como expandir a base de voluntários capacitados? Necessidade de desenvolver materiais de formação em português, tutoriais específicos para edição de recursos lexicográficos e workshops de capacitação para envolver a comunidade académica e o público interessado.

Sustentabilidade técnica

Questões sobre manutenção a longo prazo: como garantir que as ligações semânticas permanecem válidas? Como atualizar predefinições sem quebrar projetos existentes? Como arquivar e documentar decisões editoriais para referência futura?

Considerações finais

- A Wikisource em português é mais do que uma biblioteca digital: é um movimento vivo de preservação e valorização do património textual lusófono.
- O estudo de caso (Dalgado, 1913) demonstra que obras lexicográficas históricas podem ser revitalizadas usando estas ferramentas.
- Esta abordagem não só fortalece a circulação e reutilização do conhecimento, como garante a sua sustentabilidade futura:
 - ◆ uma comunidade ativa assegura atualização contínua e amplia o impacto para além da academia, envolvendo toda a sociedade na preservação da sua herança cultural e linguística.

Obrigada!

acsalgado@letras.up.pt

leonor.olivmartins222gmail.com

mleonoreis@gmail.com

carlos.silva@wikimedia.pt

ltrigo@letras.up.pt

andre.barbosa@wikimedia.pt

Ligações úteis:

[Wikisource \(pt\)](#)

[Livro de Estilo](#)

[Commons Graphic Lab](#)

[Propriedades Wikidata](#)

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P.: Centro de Linguística da Universidade do Porto – UID/00022.