O projeto *Edição Digital dos Vocabulários da Academia das Ciências de Lisboa:* o VOLP 1940

Ana Salgado[1, 2], Rute Costa[1]

[1] NOVA CLUNL, Universidade NOVA de Lisboa

[2] Academia das Ciências de Lisboa

anasalgado@campus.fcsh.unl.pt

rute.costa@fcsh.unl.pt

XXXV Encontro Nacional da Associação Portuguesa de Linguística, Universidade do Minho, Portugal

VOLP 1940

Projeto financiado por:

- Projeto Estratégico do Centro de Linguística da Universidade NOVA de Lisboa (UID/LIN/03213/2019), financiado pela Fundação para a Ciência e a Tecnologia (FCT)

- Projeto Elexis – European Lexicographic Infrastructure (Horizon 2020 – Ref. 731015)

VOLP 1940

Visão geral

1. Motivação
2. Quadro teórico
3. Apresentação do **VOLP 1940**
4. Objetivos
5. Metodologia
6. Aplicação e codificação em TEI
7. Conclusões e trabalho futuro

VOLP 1940

1. Motivação

digitalização de obras lexicográficas portuguesas de referência

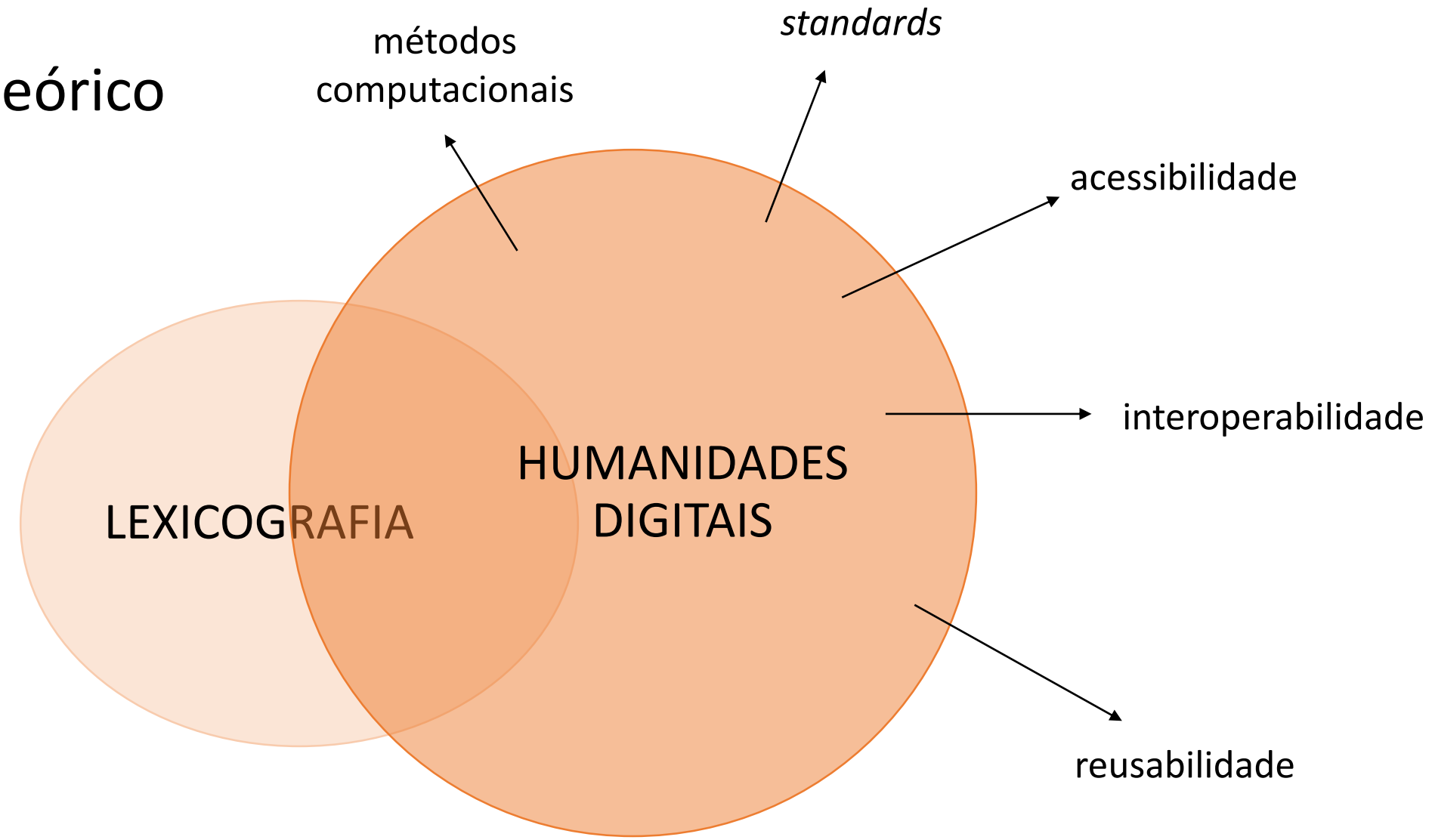
- Protocolo entre a Academia das Ciências de Lisboa (ACL), por intermédio do Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP), e a Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa (FCSH NOVA), por intermédio do Centro de Linguística da Universidade NOVA de Lisboa (NOVA CLUNL)

criação de um *corpus* lexicográfico digital que reúna todos os vocabulários da ACL

- Edições impressas dos vocabulários académicos de 1940, 1947, 1970, 2012 acessíveis a toda a comunidade
- A última versão do vocabulário já é digital (coord. Salgado, 2018, https://volp-acl.pt/)
- Disponibilização de mais recursos lexicográficos portugueses em linha

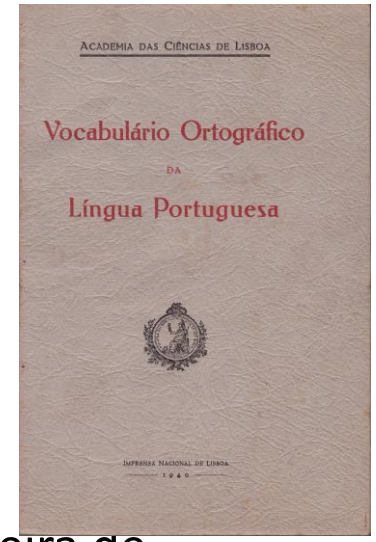
VOLP 1940

2. Quadro teórico



VOLP 1940

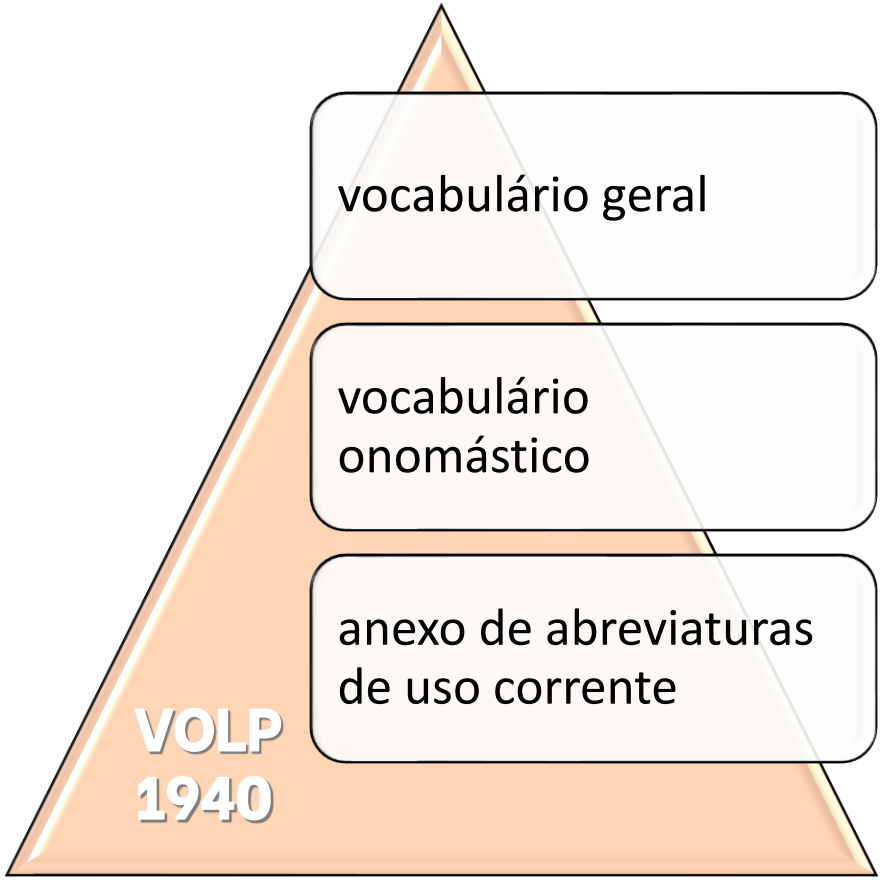
3. Apresentação do **VOLP 1940**



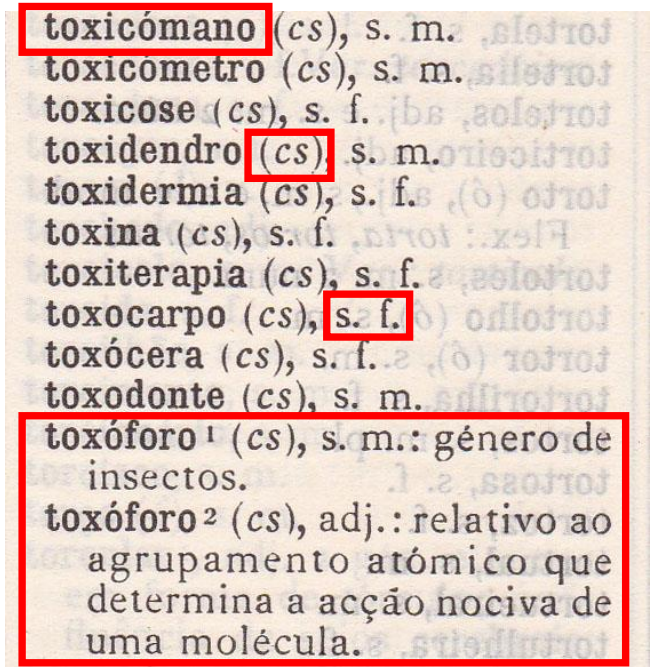
- primeiro vocabulário ortográfico com a chancela da ACL
- publicado pela Imprensa Nacional de Lisboa
- ferramenta para discutir uma nova medida ortográfica entre ACL e Academia Brasileira de Letras, que resultou na convenção ortográfica luso-brasileira de 1945 em vigor até 2011
- segue a base da reforma de 1911, recorrendo a outras duas bases: a de 1920 e o Acordo Ortográfico Luso-Brasileiro de 1931
- a nomenclatura abrange a língua portuguesa moderna, isto é, o período linguístico que decorre do século XVI até 1940
- parte fundamental do património cultural português

VOLP 1940

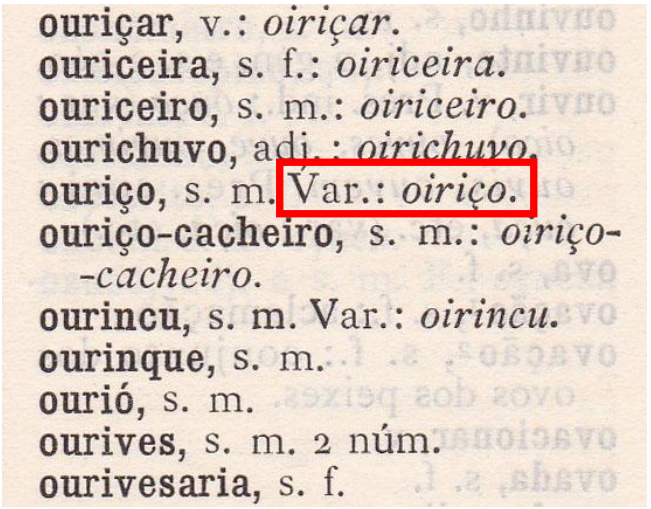
3. Apresentação do VOLP 1940



VOLP 1940



toxicómano (*cs*), s. m.
toxicómetro (*cs*), s. m.
toxicose (*cs*), s. f.
toxidendro (*cs*), s. m.
toxidermia (*cs*), s. f.
toxina (*cs*), s. f.
toxiterapia (*cs*), s. f.
toxocarpo (*cs*), s. f.
toxócera (*cs*), s. f.
toxodonte (*cs*), s. m.
toxóforo[1] (*cs*), s. m.: género de insectos.
toxóforo[2] (*cs*), adj.: relativo ao agrupamento atómico que determina a acção nociva de uma molécula.



ouriçar, v.: *oiriçar*.
ouriceira, s. f.: *oiriceira*.
ouriceiro, s. m.: *oiriceiro*.
ourichuvo, adj.: *oirichuvo*.
ouriço, s. m. Var.: *oiriço*.
ouriço-cacheiro, s. m.: *oiriço-cacheiro*.
ourincu, s. m. Var.: *oirincu*.
ourinque, s. m.
ourió, s. m.
ourives, s. m. 2 núm.
ourivesaria, s. f.

Entrada (lema)

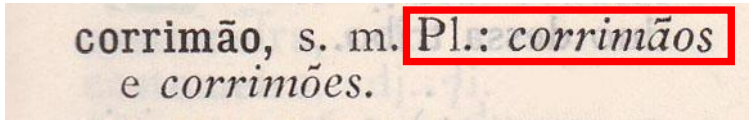
Ortoépia (apenas nas palavras de pronúncia duvidosa ou determinados timbres)

Categoria gramatical

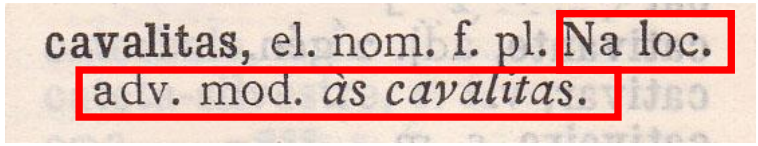
Significado

Variante ortográfica

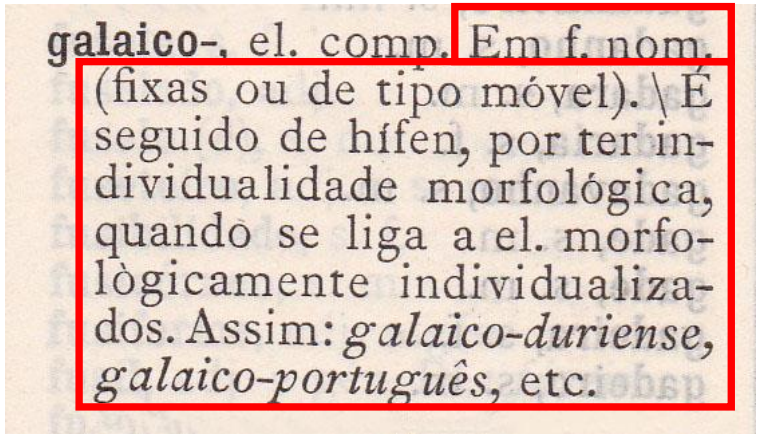
VOLP 1940


corrimão, s. m. Pl.: *corrimãos* e *corrimões.*

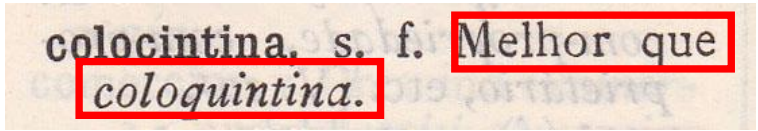
Informações morfológicas


cavalitas, el. nom. f. pl. Na loc. adv. mod. *às cavalitas.*

Informações de uso


galaico-, el. comp. Em f. nom. (fixas ou de tipo móvel). \ É seguido de hífen, por ter individualidade morfológica, quando se liga a el. morfològicamente individualizados. Assim: *galaico-duriense, galaico-português,* etc.

Regras de hifenização


colocintina, s. f. Melhor que *coloquintina.*

Notas de uso

VOLP 1940

4. Objetivos

- criar um novo recurso lexicográfico em linha
- melhorar a consistência dos metadados originais, conforme as recomendações TEI (TEI Lex-0)
- descrever a anotação linguística para um posterior enriquecimento semântico da base textual informatizada
- acrescentar novos metadados

VOLP 1940

5. Metodologia

- Digitalização da obra (OCR com qualidade)
- Correções manuais do OCR
- Análise lexicográfica minuciosa: identificação dos elementos microestruturais dos artigos
- Utilização de standards – formato XML, TEI
- Criação de uma base de dados com pesquisa avançada

Simões *et al.* (2019). LeXmart: A smart tool for lexicographers, Elex 2019 http://www.lexmart.eu/

- Enriquecimento da base de dados lexicais

VOLP 1940

6. Aplicação e codificação em TEI

A. *TEI Guidelines* para a codificação de dicionários

- A TEI – *Text Encoding Initiative* (Iniciativa de Codificação Textual, https://www.tei-c.org) – é uma norma *de facto* internacional e interdisciplinar.

- A TEI é uma referência em projetos de edição digital ou anotação de texto, desde cartas a notação musical.

- *TEI Guidelines* (https://tei-c.org/guidelines) especifica como codificar os textos e tem um capítulo (9) inteiramente dedicado a dicionários.

VOLP 1940

6. Aplicação e codificação em TEI

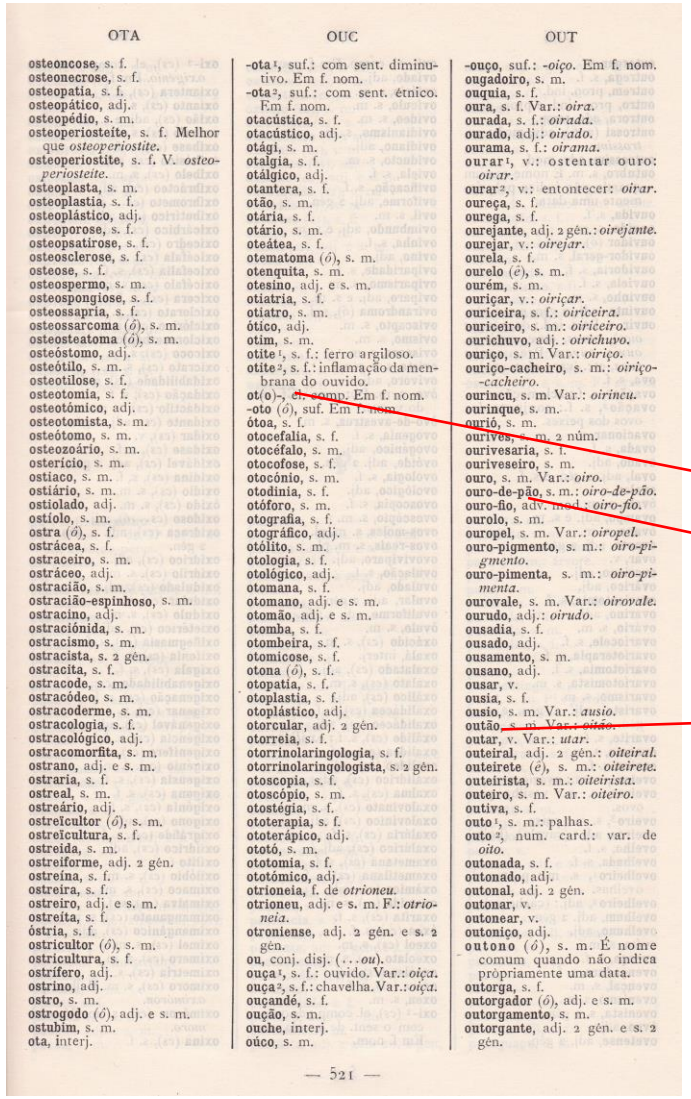
B. Aplicação de TEI Lex-0

- Certas soluções nas Diretrizes não cobrem as necessidades dos recursos lexicográficos portugueses.
- As Diretrizes (*Guidelines*) contêm várias possibilidades de codificação para os mesmos componentes.
- TEI Lex-0: um novo formato, simplificação do capítulo 9 das Directrizes.
- O esquema não é fixo (participação activa em: https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0).

Salgado *et al.* (2019). TEI Lex-0 in action, Elex 2019.

Romary & Tasovac (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources.

VOLP 1940



OTA OUC OUT

osteoncose, s. f.
osteonecrose, s. f.
osteopatia, s. f.
osteopático, adj.
osteopédio, s. m.
osteoperiosteíte, s. f. Melhor que *osteoperiostite*.
osteoperiostite, s. f. V. *osteoperiosteíte*.
osteoplasta, s. m.
osteoplastia, s. f.
osteoplástico, adj.
osteoporose, s. f.
osteopsatirose, s. f.
osteosclerose, s. f.
osteose, s. f.
osteospermo, s. m.
osteospongiose, s. f.
osteossapria, s. f.
osteossarcoma (*ó*), s. m.
osteosteatoma (*ó*), s. m.
osteóstomo, adj.
osteótilo, s. m.
osteotilose, s. f.
osteotomia, s. f.
osteotómico, adj.
osteotomista, s. m.
osteótomo, s. m.
osteozoário, s. m.
osterício, s. m.
ostiaco, s. m.
ostiário, s. m.
ostiolado, adj.
ostíolo, s. m.
ostra (*ó*), s. f.
ostrácea, s. f.
ostraceiro, s. m.
ostráceo, adj.
ostracião, s. m.
ostracião-espinhoso, s. m.
ostracino, adj.
ostraciónida, s. m.
ostracismo, s. m.
ostracista, s. 2 gén.
ostracita, s. f.
ostracode, s. m.
ostracódeo, s. m.
ostracoderme, s. m.
ostracologia, s. f.
ostracológico, adj.
ostracomorfita, s. m.
ostrano, adj. e s. m.
ostraria, s. f.
ostreal, s. m.
ostreário, adj.
ostreícultor (*ó*), s. m.
ostreícultura, s. f.
ostreida, s. m.
ostreiforme, adj. 2 gén.
ostreína, s. f.
ostreira, s. f.
ostreiro, adj. e s. m.
ostreíta, s. f.
óstria, s. f.
ostricultor (*ó*), s. m.
ostricultura, s. f.
ostrífero, adj.
ostrino, adj.
ostro, s. m.
ostrogodo (*ó*), adj. e s. m.
ostubim, s. m.
ota, interj.

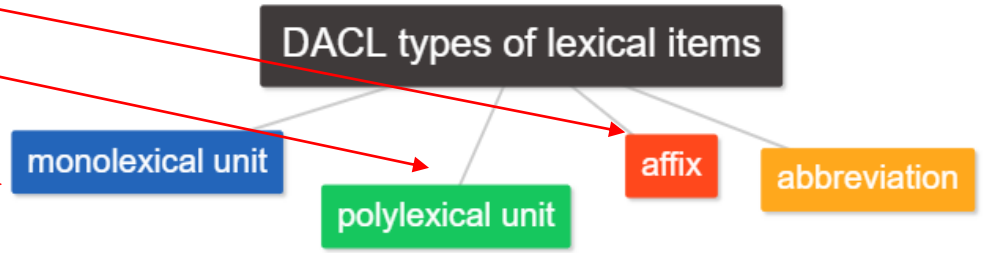
-ota[1], suf.: com sent. diminutivo. Em f. nom.
-ota[2], suf.: com sent. étnico. Em f. nom.
otacústica, s. f.
otacústico, adj.
otági, s. m.
otalgia, s. f.
otálgico, adj.
otantera, s. f.
otão, s. m.
otária, s. f.
otário, s. m.
oteátea, s. f.
otematoma (*ó*), s. m.
otenquita, s. m.
otesino, adj. e s. m.
otiatria, s. f.
otiatro, s. m.
ótico, adj.
otim, s. m.
otite[1], s. f.: ferro argiloso.
otite[2], s. f.: inflamação da membrana do ouvido.
ot(o)-, el. comp. Em f. nom.
-oto (*ó*), suf. Em f. nom.
ótoa, s. f.
otocefalia, s. f.
otocéfalo, s. m.
otocofose, s. f.
otocónio, s. m.
otodínia, s. f.
otóforo, s. m.
otografia, s. f.
otográfico, adj.
otólito, s. m.
otologia, s. f.
otológico, adj.
otomana, s. f.
otomano, adj. e s. m.
otomão, adj. e s. m.
otomba, s. f.
otombeira, s. f.
otomicose, s. f.
otona (*ó*), s. f.
otopatia, s. f.
otoplastia, s. f.
otoplástico, adj.
otorcular, adj. 2 gén.
otorreia, s. f.
otorrinolaringologia, s. f.
otorrinolaringologista, s. 2 gén.
otoscopia, s. f.
otoscópio, s. m.
otostégia, s. f.
ototerapia, s. f.
ototerápico, adj.
ototó, s. m.
ototomia, s. f.
ototómico, adj.
otrioneia, f. de *otrioneu*.
otrioneu, adj. e s. m. F.: *otrioneia*.
otroniense, adj. 2 gén. e s. 2 gén.
ou, conj. disj. (*...ou*).
ouça[1], s. f.: ouvido. Var.: *oiça*.
ouça[2], s. f.: chavelha. Var.: *oiça*.
ouçandé, s. f.
oução, s. m.
ouche, interj.
oúco, s. m.

-ouço, suf.: *-oiço*. Em f. nom.
ougadoiro, s. m.
ouquia, s. f.
oura, s. f. Var.: *oira*.
ourada, s. f.: *oirada*.
ourado, adj.: *oirado*.
ourama, s. f.: *oirama*.
ourar[1], v.: ostentar ouro: *oirar*.
ourar[2], v.: entontecer: *oirar*.
oureça, s. f.
ourega, s. f.
ourejante, adj. 2 gén.: *oirejante*.
ourejar, v.: *oirejar*.
ourela, s. f.
ourelo (*é*), s. m.
ourém, s. m.
ouriçar, v.: *oiriçar*.
ouriceira, s. f.: *oiriceira*.
ouriceiro, s. m.: *oiriceiro*.
ourichuvo, adj.: *oirichuvo*.
ouriço, s. m. Var.: *oiriço*.
ouriço-cacheiro, s. m.: *oiriço-cacheiro*.
ourincu, s. m. Var.: *oirincu*.
ourinque, s. m.
ourió, s. m.
ourives, s. m. 2 núm.
ourivesaria, s. f.
ouriveseiro, s. m.
ouro, s. m. Var.: *oiro*.
ouro-de-pão, s. m.: *oiro-de-pão*.
ouro-fio, adv. mod.: *oiro-fio*.
ourolo, s. m.
ouropel, s. m. Var.: *oiropel*.
ouro-pigmento, s. m.: *oiro-pigmento*.
ouro-pimenta, s. m.: *oiro-pimenta*.
ourovale, s. m. Var.: *oirovale*.
ourudo, adj.: *oirudo*.
ousadia, s. f.
ousado, adj.
ousamento, s. m.
ousano, adj.
ousar, v.
ousia, s. f.
ousio, s. m. Var.: *ausio*.
outão, s. m. Var.: *oitão*.
outar, v. Var.: *utar*.
outeiral, adj. 2 gén.: *oiteiral*.
outeirete (*é*), s. m.: *oiteirete*.
outeirista, s. m.: *oiteirista*.
outeiro, s. m. Var.: *oiteiro*.
outiva, s. f.
outo[1], s. m.: palhas.
outo[2], num. card.: var. de *oito*.
outonada, s. f.
outonado, adj.
outonal, adj. 2 gén.
outonar, v.
outonear, v.
outoniço, adj.
outono (*ó*), s. m. É nome comum quando não indica propriamente uma data.
outorga, s. f.
outorgador (*ó*), adj. e s. m.
outorgamento, s. m.
outorgante, adj. 2 gén. e s. 2 gén.

— 521 —

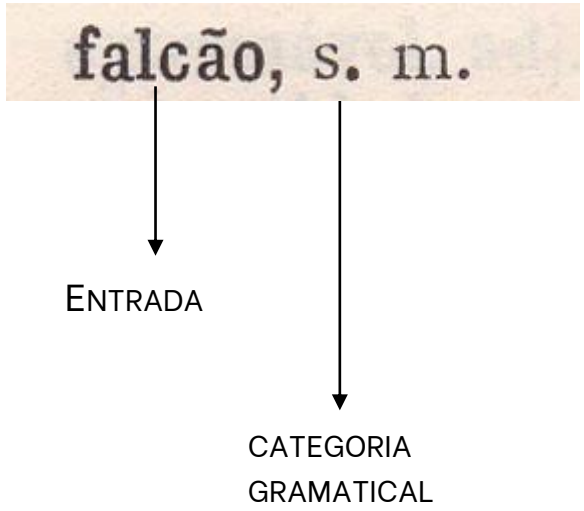
- Classificação para otimizar o processo do trabalho lexicográfico
- Controlar o vocabulário
- Aproveitamento do trabalho desenvolvido no Dicionário ACL

Salgado *et al.* (2019). TEI Lex-0 in action, Elex 2019



VOLP 1940

Estrutura básica de uma entrada



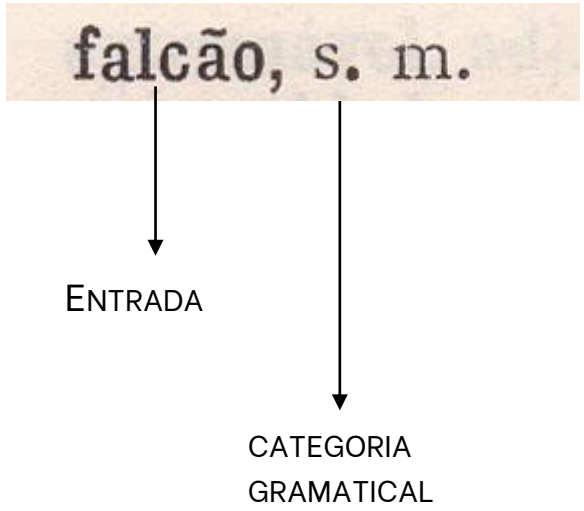

```
<entry xml:lang="pt" xml:id="falcão">
    <form type="lemma">
      <orth>falcão</orth>
    </form>
    <gramGrp>
      <gram type="lexicalConstruction" value="monolexical"></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
    </gramGrp>
</entry>
```

entry
- @xml:id
- @xml:lang – code (BCP 47)

gram type element
- categoria grammatical e outras especificações
- atributo @norm attribute – Universal Dependencies Part-of-Speech: https://universaldependencies.org/u/pos/

VOLP 1940

Estrutura básica de uma entrada



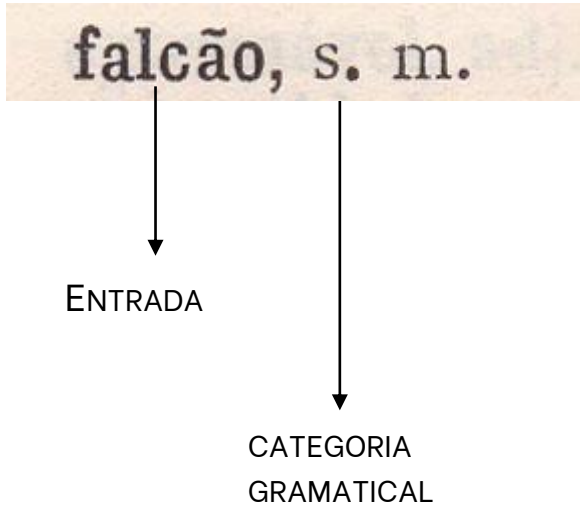

```
<entry xml:lang="pt" xml:id="falcão">
    <form type="lemma">
        <orth>falcão</orth>
    </form>
    <gramGrp>
        <gram type="lexicalConstruction" value="monolexical"></gram>
        <gram type="pos" norm="NOUN">s.</gram>
        <gram type="gen">m.</gram>
    </gramGrp>
</entry>
```

entry
- @xml:id
- @xml:lang – code (BCP 47)

gram type element
- categoria grammatical e outras especificações
- atributo @norm attribute – **Universal Dependencies Part-of-Speech**: https://universaldependencies.org/u/pos/

VOLP 1940

Estrutura básica de uma entrada



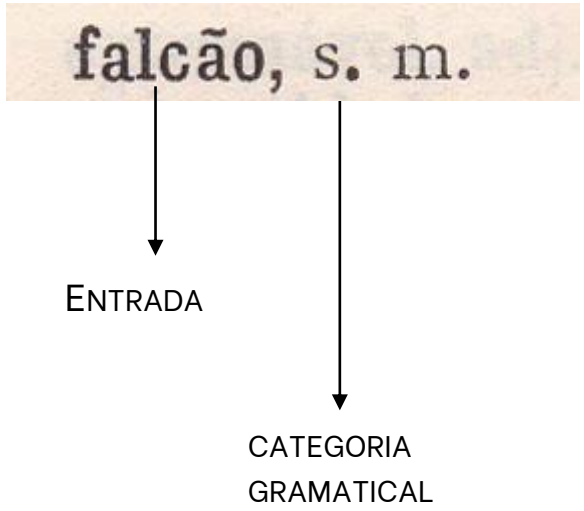

```
<entry xml:lang="pt" xml:id="falcão">
    <form type="lemma">
      <orth>falcão</orth>
    </form>
    <gramGrp>
      <gram type="lexicalConstruction" value="monolexical"></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
    </gramGrp>
</entry>
```

entry
- @xml:id
- @xml:lang – code (BCP 47)

gram type element
- categoria grammatical e outras especificações
- atributo @norm attribute – Universal Dependencies Part-of-Speech: https://universaldependencies.org/u/pos/

VOLP 1940

Estrutura básica de uma entrada

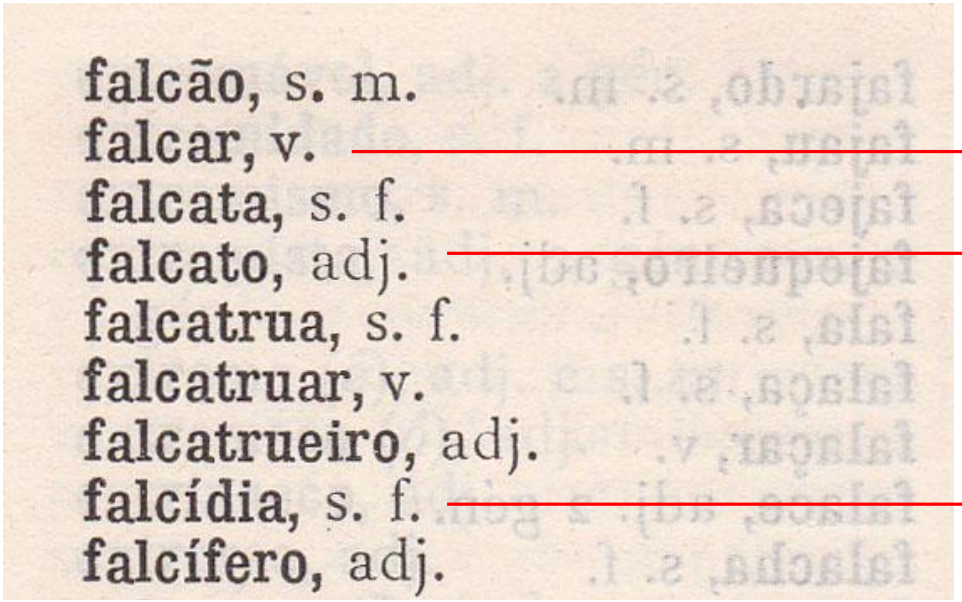



```
<entry xml:lang="pt" xml:id="falcão">
    <form type="lemma">
      <orth>falcão</orth>
    </form>
    <gramGrp>
      <gram type="lexicalConstruction" value="monolexical"></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
    </gramGrp>
</entry>
```

entry
- @xml:id
- @xml:lang – code (BCP 47)

gram type element
- categoria grammatical e outras especificações
- atributo @norm attribute – Universal Dependencies Part-of-Speech: https://universaldependencies.org/u/pos/

VOLP 1940


falcão, s. m.
falcar, v.
falcata, s. f.
falcato, adj.
falcatrua, s. f.
falcatruar, v.
falcatrueiro, adj.
falcídia, s. f.
falcífero, adj.

VERB

ADJECTIVE

NOUN

elexis european lexicographic infrastructure

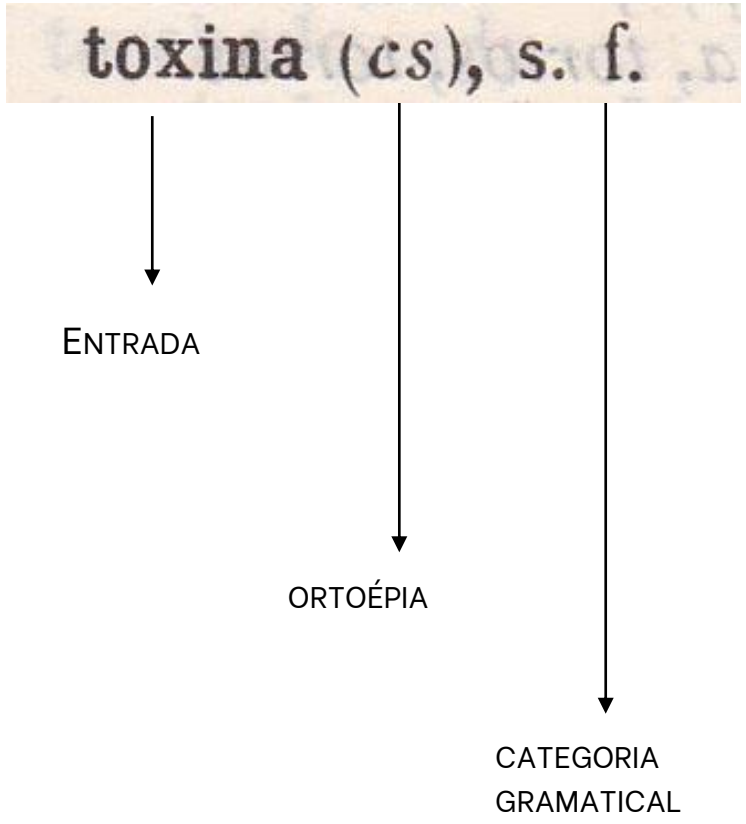
Universal POS tags

These tags mark the core part-of-speech categories. To distinguish additional lexical and grammatical properties of words, use the universal features.

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

https://universaldependencies.org/u/pos/

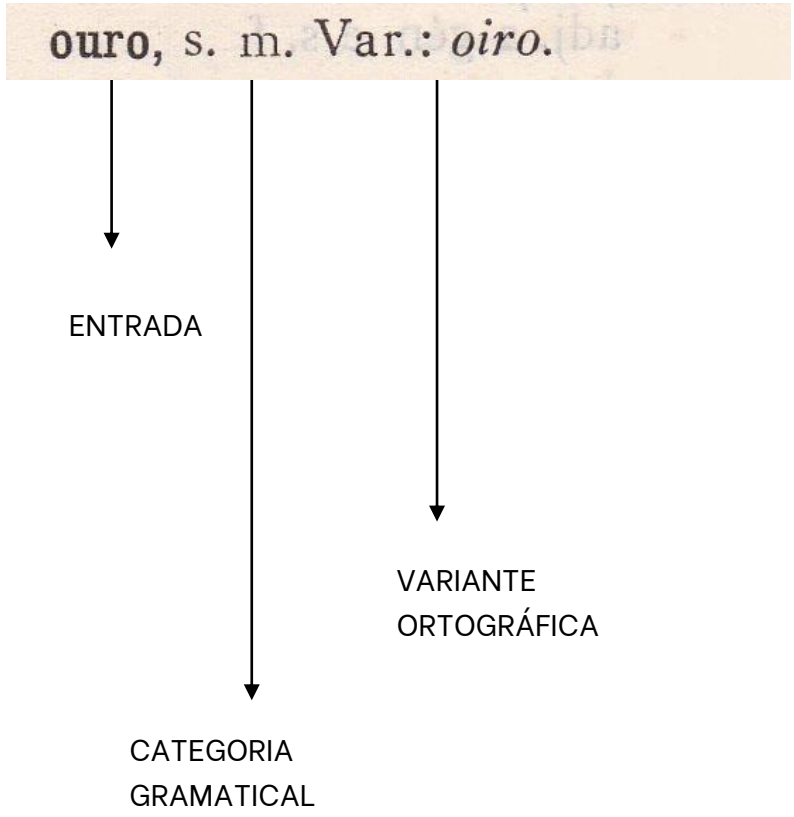
VOLP 1940




```
<entry xml:lang="pt" xml:id="toxina">
    <form type="lemma">
      <orth>toxina</orth>
      <pron>cs</pron>
    </form>
    <gramGrp>
      <gram type="lexicalConstruction" value="monolexical"></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
</entry>
```

VOLP 1940

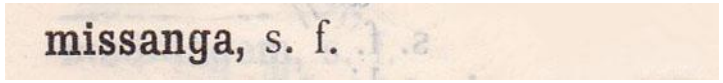
Enriquecimento da base textual




```
<entry xml:lang="pt" xml:id="ouro">
    <form type="lemma">
      <orth>ouro</orth>
    </form>
  </form>
  <form type="variant" xml:id="oiro" xml:lang="pt">
  <orth>oiro</orth>
  </form>
  <gramGrp>
      <gram type="lexicalConstruction" value="monolexical"></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
</entry>
```

VOLP 1940

Enriquecimento da base textual


missanga, s. f.

```
<entry xml:lang="pt" xml:id="missanga">
    <form type="lemma">
      <orth>missanga</orth>
    </form>
   </form>
   <form type="variant" xml:id="miçanga" xml:lang="pt">
     <usg type="geo"> <placeName>Brasil</placeName>
     <orth>miçanga</orth>
   </form>
   <gramGrp>
      <gram type= "lexicalConstruction" value=monolexical></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
</entry>
```

VOLP 1940

Enriquecimento da base textual

ouriço-cacheiro, s. m.: *oiriço-cacheiro*.

colibri, s. m.

colibri [kɔlibrí]. *s. m.* (Do caribe *kolibris*, pelo cast. *colibrí*). *Zool.* Designação vulgar de várias aves da família dos troquilídeos (*Trochilus*, Lin.), de tamanho reduzido, plumagem de cores vivas e brilhantes, voo muito veloz, frequentes na América tropical, também conhecidas por *beija-flor*, *chupa-flor*, *chupa-mel* e *pica-flor*.

colibri

nome masculino

DLPC-ACL, 2001

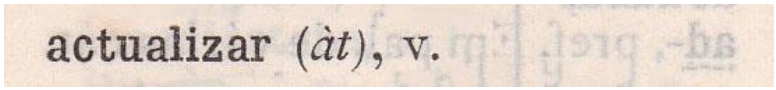
VOLP-ACL, 2018

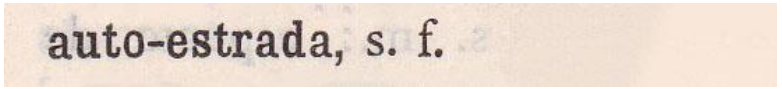
```
<entry xml:id="ouriço-cacheiro" xml:lang="pt">
<form type="lemma">
<orth>ouriço-cacheiro</orth>
<orth type="variant">oiriço-cacheiro</orth>
</form>
<gram type="pos" ud:norm="NOUN">n.</gram>
<gram type="gen">m.</gram>
<sense>
<usg type="domain">Zool.</usg>
<form type="inflected">
<gram type="number">Pl.</gram></gramGrp>
<orth>ouriços-cacheiros</orth>
<orth type="variant">oiriços-cacheiros</orth>
</form>
</sense>
</entry>

<entry xml:id="colibri" xml:lang="pt">
<form type="lemma">
<orth>colibri</orth>
</form>
<!--etc. -->
<sense>
 <usg type="domain">Zool.</usg>
 <xr type="synonymy"><ref type="sense">beija-flor</ref></xr>
 <xr type="synonymy"><ref type="sense">chupa-flor</ref></xr>
 <xr type="synonymy"><ref type="sense">chupa-mel</ref></xr>
 <xr type="synonymy"><ref type="sense">pica-flor</ref></xr>
 </sense>
 </entry>
```

VOLP 1940

Enriquecimento da base textual


actualizar (*àt*), v.


auto-estrada, s. f.

actualizar > atualizar

```
<entry type="simple" xml:id="actualizar" xml:lang="pt">
<form type="lemma">
<orth>actualizar</orth>
</form>
<form type="variant">
<orth notBefore="2011" xml:lang="pt-PT">actualizar</orth>
<usg type="time">Acordo Ortográfico de 1990</usg>
</form>
</entry>
```

auto-estrada > autoestrada

```
<entry type="simple" xml:id="auto-estrada" xml:lang="pt">
<form type="lemma">
<orth>auto-estrada</orth>
</form>
<form type="variant">
<orth notBefore="2011" xml:lang="pt-PT">autoestrada</orth>
<usg type="time">Acordo Ortográfico de 1990</usg>
</form>
</entry>
```

VOLP 1940

7. Conclusões e trabalho futuro

- disponibilizar um novo recurso lexicográfico português em linha que reúna as versões impressas dos vocabulários do ACL (1940, 1947, 1970, 2012) e melhore as múltiplas funcionalidades de pesquisa, como fonte de pesquisa científica e património cultural

- estabelecer *links* para recursos digitais externos e para as entradas do dicionário da ACL sob revisão

- a codificação é mais pormenorizada em TEI Lex-0, mas mais estrutural e rigorosa, permitindo que as máquinas processem melhor essas informações

Obrigada pela atenção!

Ana Salgado & Rute Costa

anasalgado@campus.fcsh.unl.pt

rute.costa@fcsh.unl.pt