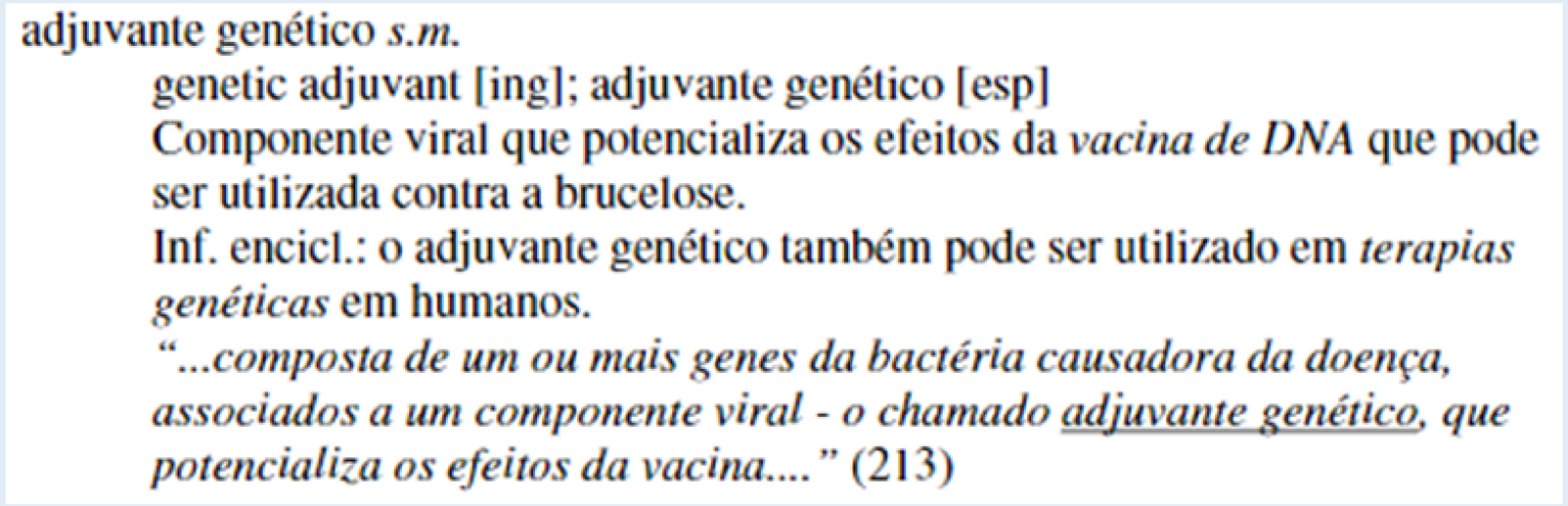
# Processamento de Linguagem Natural

## Trabalho Prático 1

Dinis Mesquita  PG57810
Flávio Ribeiro PG52290
Sofia Corriea PG56152

# Glossário de Neologismos Terminológicos da Saúde Humana

Apresenta uma estrutura de tese, fator levado em consideração no processamento do mesmo. Consta com 306 conceitos e cada um dos mesmos apresenta das traduções inglesas e espanholas, o significado, significado enciclopédico e em alguns casos um contexto textual.



adjuvante genético *s.m.*

genetic adjuvant [ing]; adjuvante genético [esp]

Componente viral que potencializa os efeitos da *vacina de DNA* que pode ser utilizada contra a brucelose.

Inf. enciclo.: o adjuvante genético também pode ser utilizado em *terapias genéticas* em humanos.

*"...composta de um ou mais genes da bactéria causadora da doença, associados a um componente viral - o chamado adjuvante genético, que potencializa os efeitos da vacina...."* (213)

Figura 1. Exemplo estrutural do Glossário de Neologismos Terminológicos da Saúde Humana

# Tratamento do Glossário de Neologismos

O passo inicial para a limpeza do documento revolveu-se em retirar toda a informação irrelevante ao nosso objetivo. Neste caso é tanto a informação que se encontra antes e depois do intervalo dos conceitos e os respetivos parâmetros como os elementos XML, Figura 2 e Figura 3.

```
<page number="1" position="absolute" top="0" left="0" height="1263" width="892">
    <fontspec id="0" size="18" family="GAAIBX+Times" color="#000000"/>
    <fontspec id="1" size="18" family="YSWOZB+Times" color="#000000"/>
<text top="128" left="310" width="320" height="21" font="0"><b>AURI CLAUDIONEI MATOS FRÜBEL </b></text>
<text top="148" left="468" width="4" height="21" font="0"><b> </b></text>
<text top="169" left="468" width="4" height="21" font="0"><b> </b></text>
<text top="190" left="468" width="4" height="21" font="0"><b> </b></text>
<text top="210" left="468" width="4" height="21" font="0"><b> </b></text>
```

Figura 2. Excerto do ficheiro XML

```
@abeta
» abeta [ing]; abeta [esp]
*Proteína que pode ser encontrada em todos os tipos de células do
organismo humano. Ao acumular-se excessivamente no córtex cerebral
do ser humano pode contribuir para o aceleramento do
mal de alzheimer .
£“... Pesquisadores alemães da Universidade de Bonn ajudaram a entender
como a proteína abeta se acumula no córtex cerebral de portadores do
mal de Alzheimer...”
```

Figura 3. Ficheiro após a limpeza e com tags

# Tratamento do Glossário de Neologismos

Por fim, foi criado um script para armazenar as informações relevantes deste glossário em formato JSON, Figura 4. Realizou-se o split a partir da tag @, de forma a separar cada conceito e o respetivo conteúdo do mesmo. A partir daí extraiu-se cada parâmetro com base na sua tag.


```
[
    {
        "conceito": "abeta",
        "tradução_eng": "abeta",
        "tradução_esp": "abeta",
        "significado": "Proteína  que  pode  ser  encontrada  em  todos  os  tipos  de  células  do
        "contexto": "... Pesquisadores alemães da Universidade de Bonn ajudaram a entender  como  a
    },
    {
        "conceito": "ação vasoconstritora",
        "tradução_eng": "vasoconstriction",
        "tradução_esp": "acción vasoconstritora",
        "significado": "Redução do diâmetro das veias artérias do organismo humano, o que  implica na
        "contexto": "...descobriram como atuam diferentes versões dos genes que controlam  a produção
    },
```


Figura 4. Ficheiro após a limpeza e com tags

# Tratamento do Glossário de Neologismos

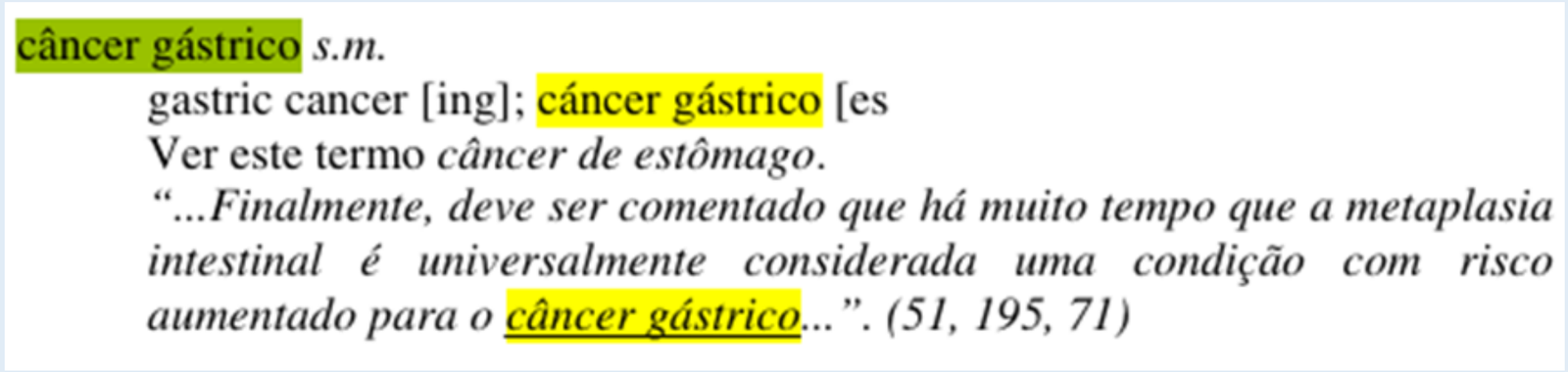
No processo de extrair os dados e verificar se toda a informação foi mantida, verificou-se a existência de falhas, relativamente às traduções espanholas. Havendo dois momentos onde estas traduções não estavam presentes, Figura 6.

```
10  # Verificar os campos sem "tradução_esp"
11  conceitos_sem_traducao_esp = [
12      conceito["conceito"] for conceito in conceitos
13      if "tradução_esp" not in conceito or not conceito["tradução_esp"].strip()
14  ]
15
16  # Resultado
17  if conceitos_sem_traducao_esp:
18      print("Conceitos sem tradução_esp:")
19      for c in conceitos_sem_traducao_esp:
20          print(c)
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

PS C:\Users\flavi\Desktop\TP1 PLN atualizado_Flavio> & C:/Users/flavi/AppData/Local/Programs/
Conceitos sem tradução_esp:
câncer gástrico
encefalopatia espongiforme
PS C:\Users\flavi\Desktop\TP1 PLN atualizado_Flavio>
```

Figura 5. Conceitos sem tradução espanhola

# Tratamento do Glossário de Neologismos



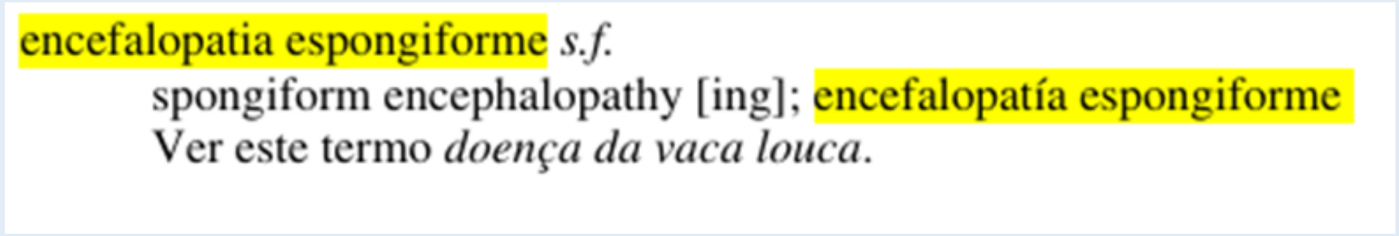
câncer gástrico *s.m.*

gastric cancer [ing]; cáncer gástrico [es

Ver este termo *câncer de estômago*.

*"...Finalmente, deve ser comentado que há muito tempo que a metaplasia intestinal é universalmente considerada uma condição com risco aumentado para o câncer gástrico...". (51, 195, 71)*

Figura 6. Conceito com tag incompleta



encefalopatia espongiforme *s.f.*

spongiform encephalopathy [ing]; encefalopatía espongiforme

Ver este termo *doença da vaca louca*.

Figura 7. Conceito com tag incompleta Conceito sem tag

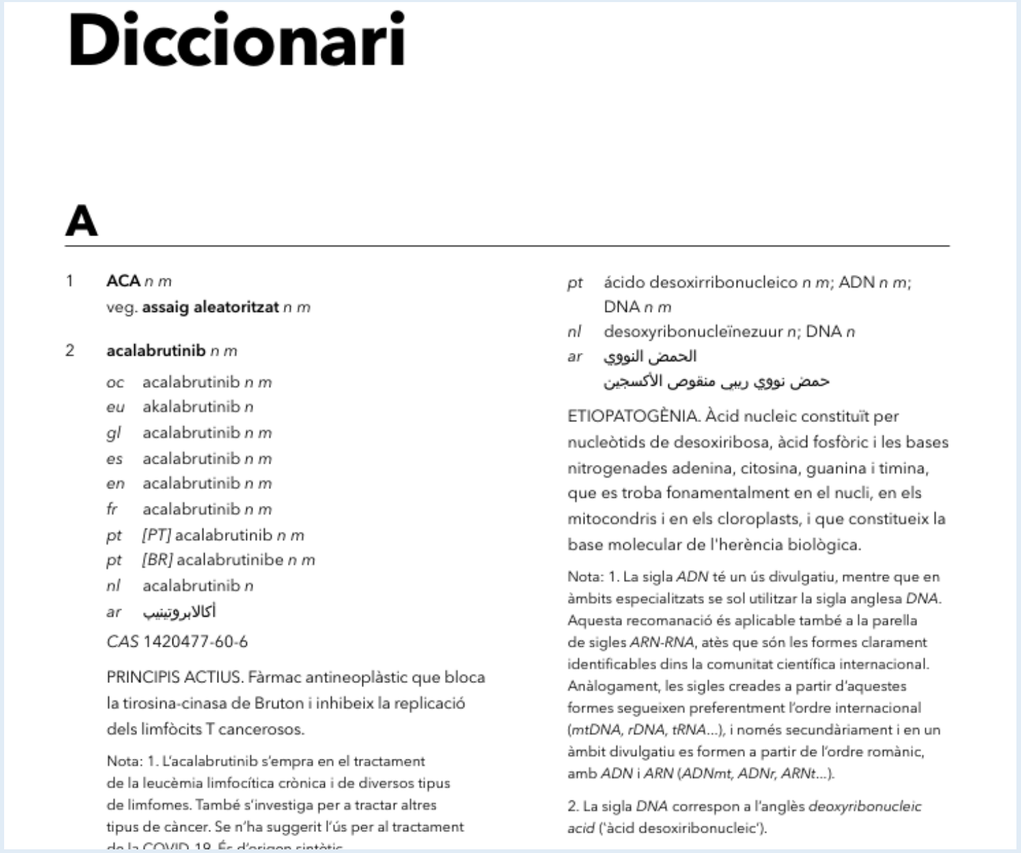
# Diccionari multilingüe de la COVID-19

- Escrito em catalão
- Maioritariamente organizado em duas colunas

- Duas partes de interesse:
  - Dicionário
  - Índices

# Diccionari multilingüe de la COVID-19

## Dicionário



Diccionari

A

1 **ACA** *n m*
veg. **assaig aleatoritzat** *n m*

2 **acalabrutinib** *n m*

*oc* acalabrutinib *n m*
*eu* akalabrutinib *n*
*gl* acalabrutinib *n m*
*es* acalabrutinib *n m*
*en* acalabrutinib *n m*
*fr* acalabrutinib *n m*
*pt* *[PT]* acalabrutinib *n m*
*pt* *[BR]* acalabrutinibe *n m*
*nl* acalabrutinib *n*
*ar* أكالابروتينيب

*CAS* 1420477-60-6

PRINCIPIS ACTIUS. Fàrmac antineoplàstic que bloca la tirosina-cinasa de Bruton i inhibeix la replicació dels limfòcits T cancerosos.

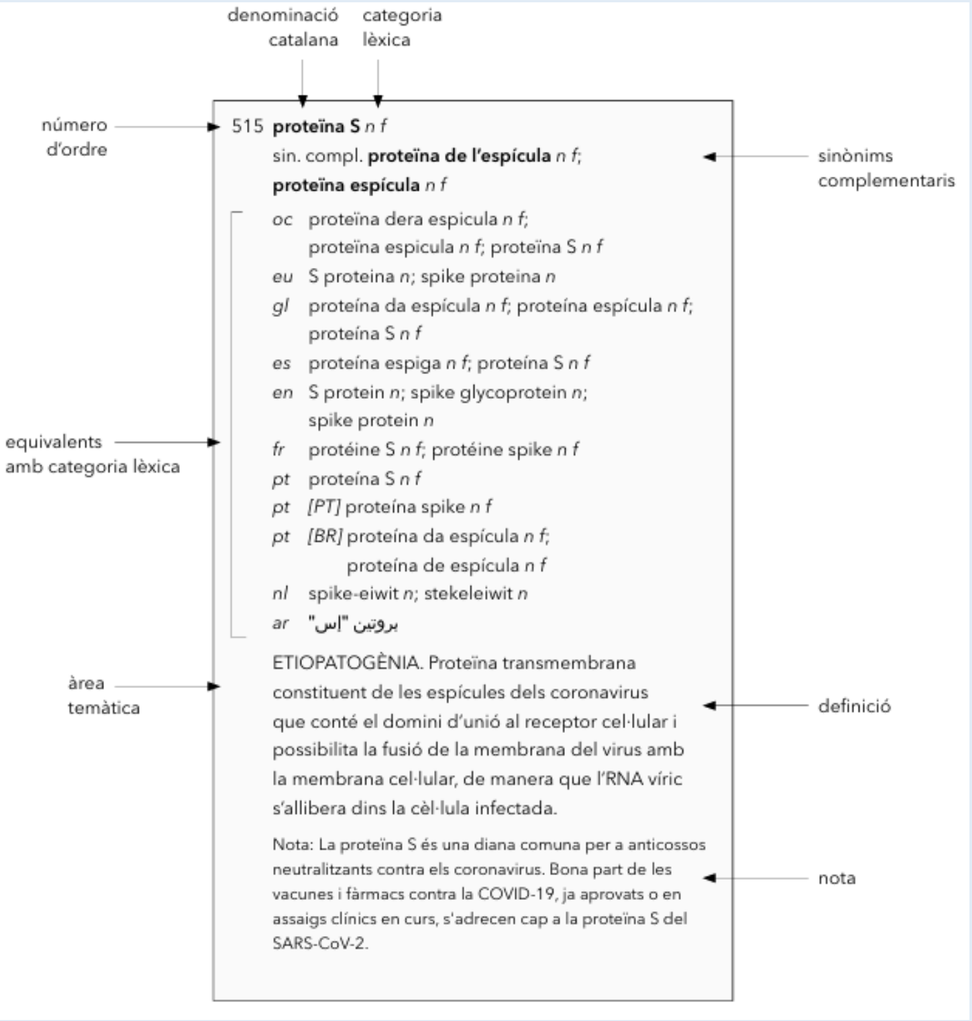
Nota: 1. L'acalabrutinib s'empra en el tractament de la leucèmia limfocítica crònica i de diversos tipus de limfomes. També s'investiga per a tractar altres tipus de càncer. Se n'ha suggerit l'ús per al tractament de la COVID-19. És d'origen sintètic.

*pt* ácido desoxirribonucleico *n m*; ADN *n m*; DNA *n m*
*nl* desoxyribonucleïnezuur *n*; DNA *n*
*ar* الحمض النووي
حمض نووي ريبي منقوص الأكسجين

ETIOPATOGÈNIA. Àcid nucleic constituït per nucleòtids de desoxiribosa, àcid fosfòric i les bases nitrogenades adenina, citosina, guanina i timina, que es troba fonamentalment en el nucli, en els mitocondris i en els cloroplasts, i que constitueix la base molecular de l'herència biològica.

Nota: 1. La sigla *ADN* té un ús divulgatiu, mentre que en àmbits especialitzats se sol utilitzar la sigla anglesa *DNA*. Aquesta recomanació és aplicable també a la parella de sigles *ARN-RNA*, atès que són les formes clarament identificables dins la comunitat científica internacional. Anàlogament, les sigles creades a partir d'aquestes formes segueixen preferentment l'ordre internacional (*mtDNA, rDNA, tRNA*...), i només secundàriament i en un àmbit divulgatiu es formen a partir de l'ordre romànic, amb *ADN* i *ARN* (*ADNmt, ADNr, ARNt*...).

2. La sigla *DNA* correspon a l'anglès *deoxyribonucleic acid* ('àcid desoxiribonucleic').

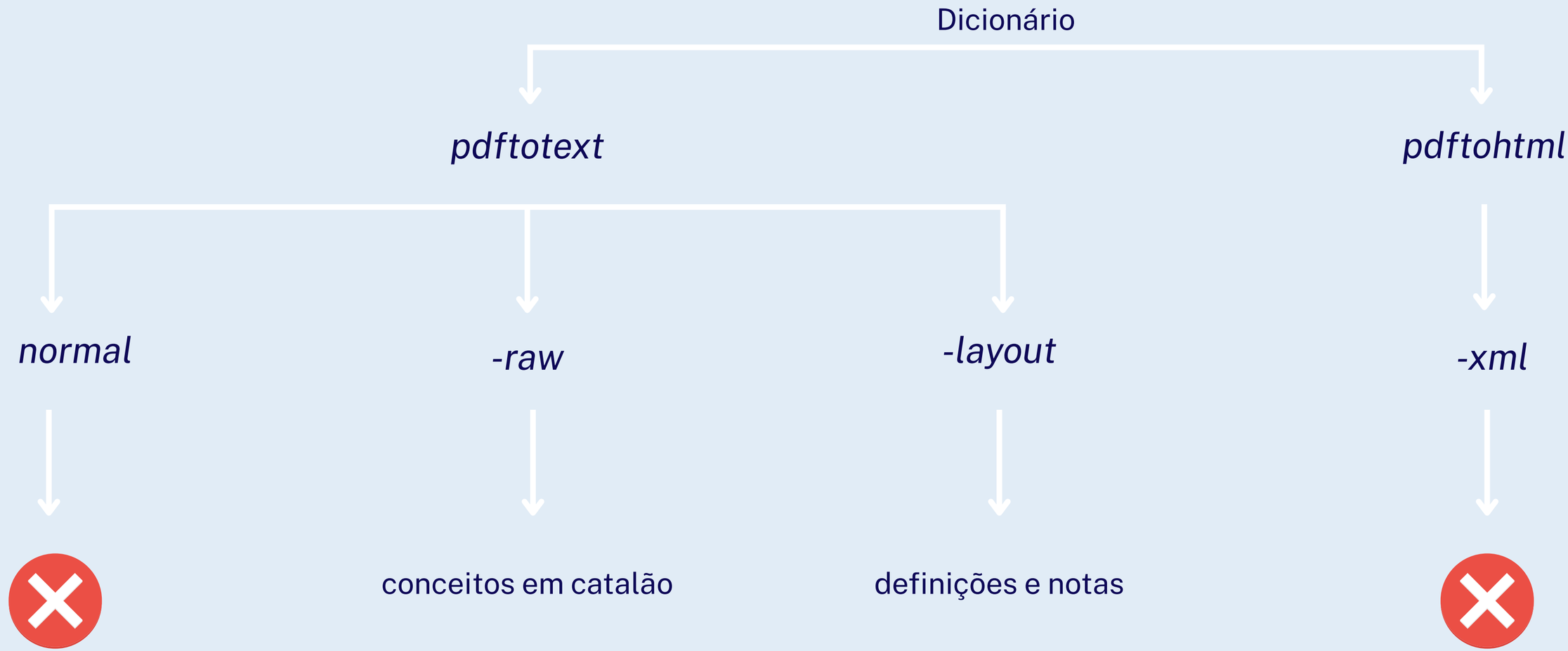


# Diccionari multilingüe de la COVID-19

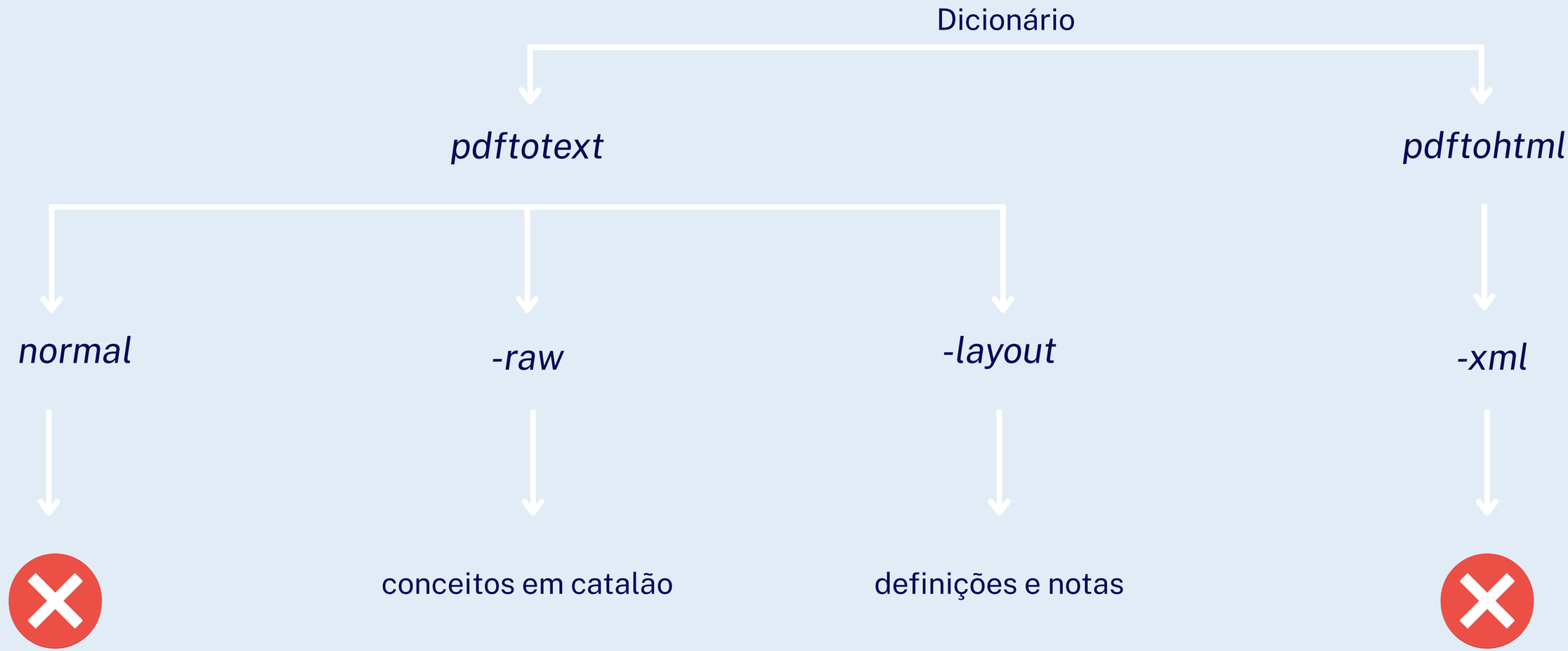
## índice

## Índex português

# Tratamento do Diccionari multilingüe de la COVID-19



# Tratamento do Diccionari multilingüe de la COVID-19



# Tratamento do Diccionari multilingüe de la COVID-19

Índices



*pdftotext*



áreas temáticas e traduções

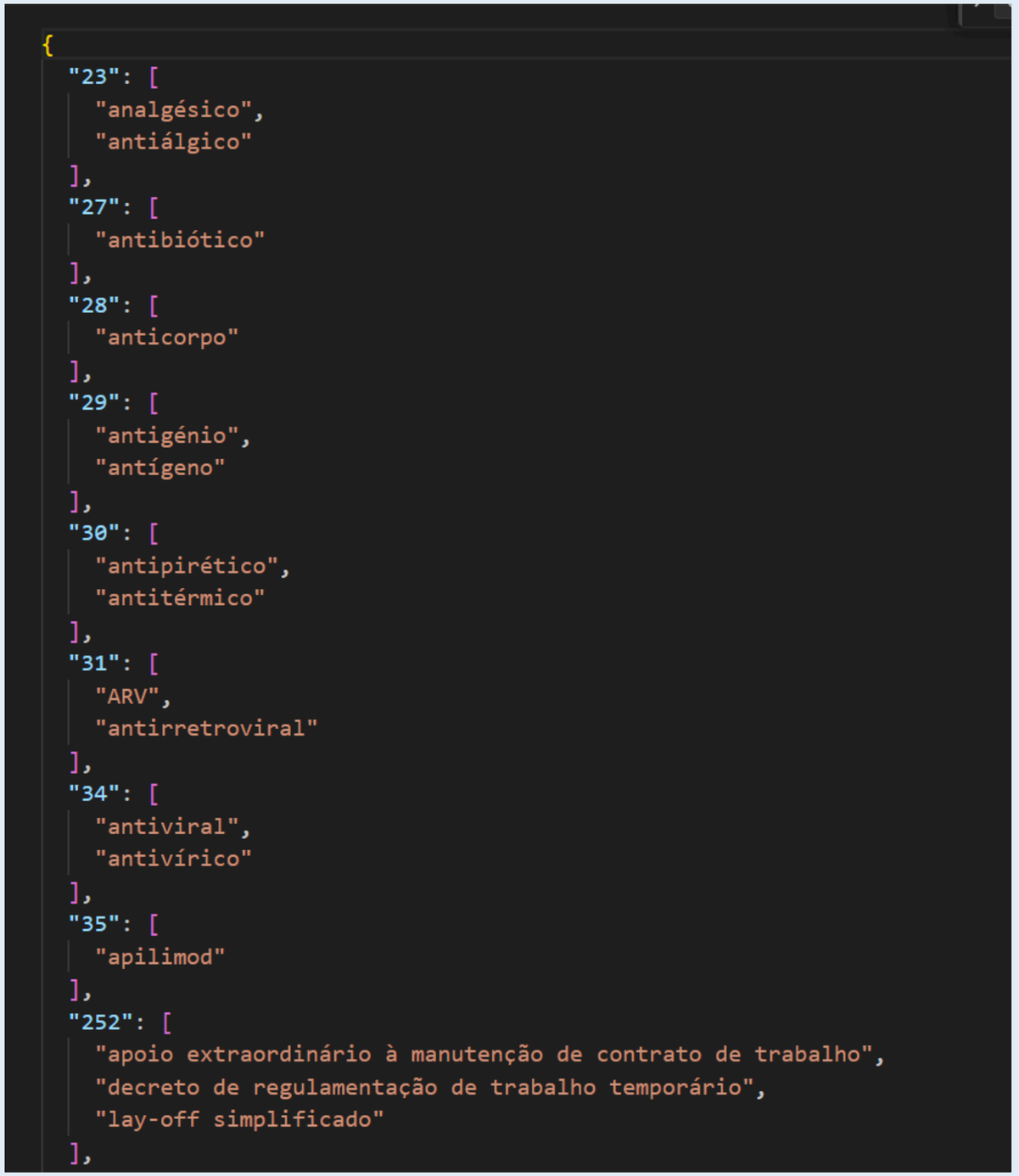
# Tratamento do Diccionari multilingüe de la COVID-19

Extrair índices

```
254
antiálgico, 23
antibiótico, 27
anticorpo, 28
antigénio, 29
antígeno, 29
antipirético, 30
antirretroviral, 31
antitérmico, 30
antiviral, 34
antivírico, 34
apilimod, 35
apoio extraordinário à manutenção de contrato de trabalho, 252
```

- retira cabeçalhos e e números de pagina;
- faz tratamento para conceitos com mais de uma linha
- separa chave/valor  por vírgulas
- para índices de línguas estrangeiras, filtra as entradas que têm equivalente em português

# Tratamento do Diccionari multilingüe de la COVID-19




```
{
  "23": [
    "analgésico",
    "antiálgico"
  ],
  "27": [
    "antibiótico"
  ],
  "28": [
    "anticorpo"
  ],
  "29": [
    "antigénio",
    "antígeno"
  ],
  "30": [
    "antipirético",
    "antitérmico"
  ],
  "31": [
    "ARV",
    "antirretroviral"
  ],
  "34": [
    "antiviral",
    "antivírico"
  ],
  "35": [
    "apilimod"
  ],
  "252": [
    "apoio extraordinário à manutenção de contrato de trabalho",
    "decreto de regulamentação de trabalho temporário",
    "lay-off simplificado"
  ],
```

# Tratamento do Diccionari multilingüe de la COVID-19

Extrair áreas temáticas:

```
183
QUADERNS 50 DICCIONARI MULTILINGÜE DE LA COVID-19
agrupament, 13
aplanar la corba, 36
bombolla, 71
bombolla ampliada, 72
bombolla de convivència, 73
brot epidèmic, 83
cadena de transmissió, 85
cadena epidemiològica, 86
cas, 98
cas confirmat, 99
cas descartat, 100
```

- separa por cada área, criando documentos txt designados a cada área
- retira cabeçalhos e e números de pagina;
- faz tratamento para conceitos com mais de uma linha
- separa por vírgulas
- para índices de línguas estrangeiras, filtra as entradas que têm equivalente em português

# Tratamento do Diccionari multilingüe de la COVID-19

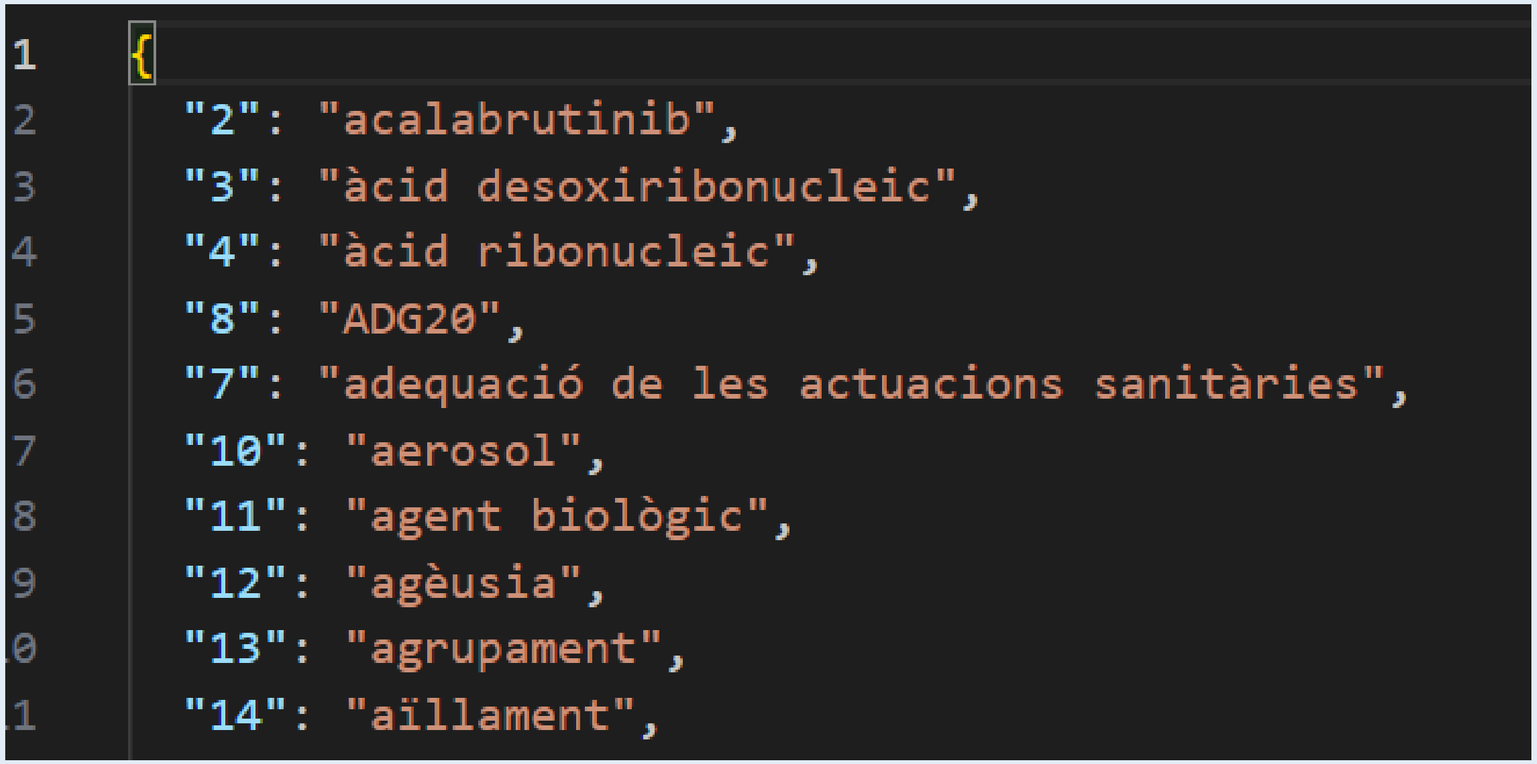
Extrair conceitos em catalão:

```python
if re.match(r'^\d+\s+', linha):
    linha = '@' + linha
```

```
2. La sigla RNA correspon a l'anglès ribonucleic acid
('àcid ribonucleic').
@5 Ad5-nCoV n f
veg. vacuna CanSino n f
@6 adequació de l'esforç terapèutic n f
veg. adequació de les actuacions sanitàries n f
@7 adequació de les actuacions sanitàries n f
sin. compl. adequació de l'esforç terapèutic n f;
limitació de l'esforç terapèutic n f
sigla LET n f
```

(com tratamento de dois casos particulares)

# Tratamento do Diccionari multilingüe de la COVID-19



```
{
  "2": "acalabrutinib",
  "3": "àcid desoxiribonucleic",
  "4": "àcid ribonucleic",
  "8": "ADG20",
  "7": "adequació de les actuacions sanitàries",
  "10": "aerosol",
  "11": "agent biològic",
  "12": "agèusia",
  "13": "agrupament",
  "14": "aïllament",
```

# Tratamento do Diccionari multilingüe de la COVID-19

Extrair definições:

**txt -layout**

(única opção que conserva ordem das linha)

**Pré-processamento para dividir cada linha em duas colunas numa dada posição**

**Marcação de pontos de interesse no texto organizado**

# Tratamento do Diccionari multilingüe de la COVID-19

Extrair definições:

marcar entrada com “«” :

```
r'(?<!-)\b(\d{1,3})(\s+\w+.*?\s(?:pl|n|m|f|adj|sigla|\||tr))\s*$',
```

marcar fim de área temática com “@” :

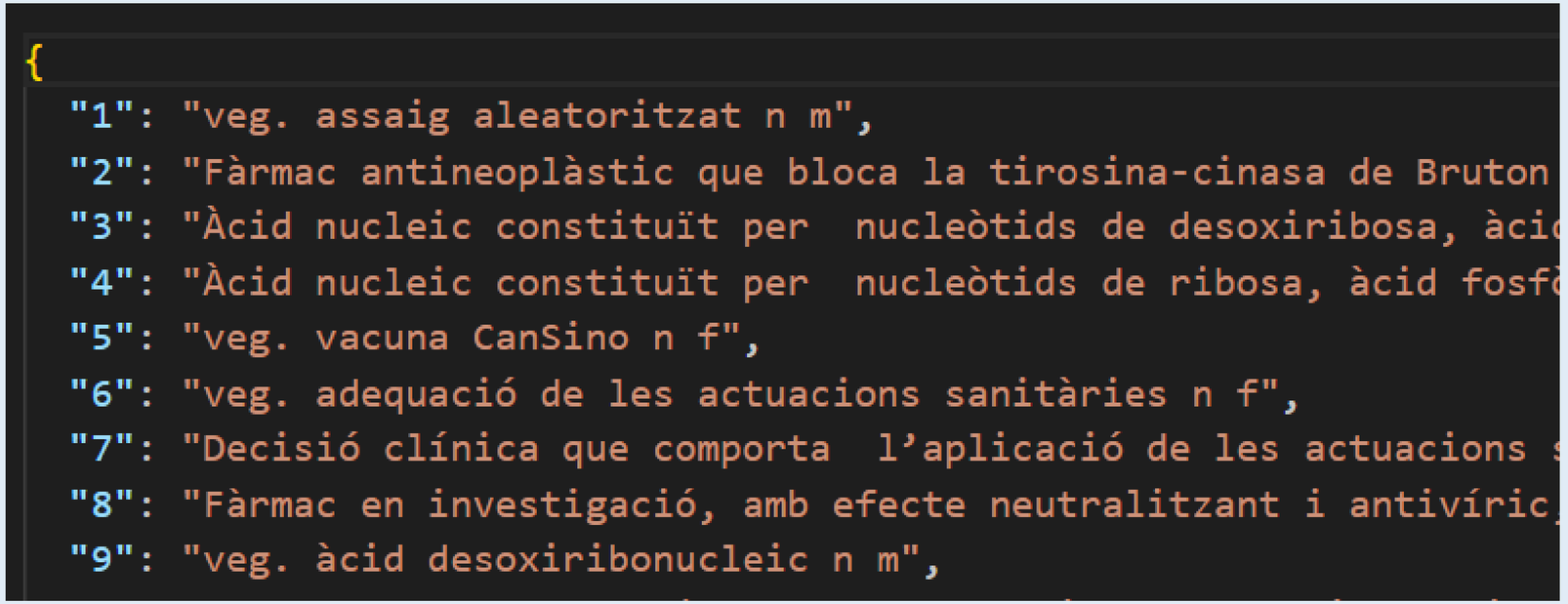
```
r'\b(CONCEPTES GENERALS|EPIDEMIOLOGIA|ETIOPATOGÈNIA|DIAGNÒSTIC|CLÍNICA|PREVENCIÓ|TRACTAMENT|PRINCIPIS ACTIUS|ENTORN SOCIAL)\.(?!@)',
```

# Tratamento do Diccionari multilingüe de la COVID-19

Extrair definições:

```
PRINCIPIS ACTIUS.@ Fàrmac antineoplàstic que bloca
la tirosina-cinasa de Bruton i inhibeix la replicac
dels limfòcits  cancerosos.

Nota: 1. ’acalabrutinib s’empra en el tractament

de la leucèmia limfocítica crònica i de diversos ti
de limfomes. També s’investiga per a tractar altres
tipus de càncer. Se n’ha suggerit l’ús per al tract
de la COVID-19. És d’origen sintètic.

2. La denominació acalabrutinib és la forma catalan
corresponent a la DCI.



«3   àcid desoxiribonucleic n m
```

# Tratamento do Diccionari multilingüe de la COVID-19



```
{
  "1": "veg. assaig aleatoritzat n m",
  "2": "Fàrmac antineoplàstic que bloca la tirosina-cinasa de Bruton
  "3": "Àcid nucleic constituït per  nucleòtids de desoxiribosa, àci
  "4": "Àcid nucleic constituït per  nucleòtids de ribosa, àcid fosf
  "5": "veg. vacuna CanSino n f",
  "6": "veg. adequació de les actuacions sanitàries n f",
  "7": "Decisió clínica que comporta  l'aplicació de les actuacions
  "8": "Fàrmac en investigació, amb efecte neutralitzant i antivíric,
  "9": "veg. àcid desoxiribonucleic n m",
```

(tendo em conta palavras com “veg.”)

# Tratamento do Diccionari multilingüe de la COVID-19

Construção do dicionário final

**Número identificativo**

**Filtrar entradas com equivalente em português, a partir do índice PT**

**Dicionário com toda a informação**

```
{
  "conceito": "eculizumab",
  "sinónimos pt": [
    "eculizumabe"
  ],
  "ar": [
    "إيكوليزوماب"
  ],
  "ca": "eculizumab",
  "en": [
    "eculizumab"
  ],
  "es": [
    "eculizumab"
  ],
  "eu": [
    "ekulizumab"
  ],
  "fr": [
    "éculizumab"
  ],
  "gl": [
    "eculizumab"
  ],
  "nl": [
    "eculizumab"
  ],
  "oc": [
    "eculizumab"
  ],
  "área médica": "Tractament",
  "definicao catalã": "Fàrmac immunomodulador selectiu que s’uneix a
},
```

# Juntar Diccionari e Glossário

Procura em todos os sinónimos pt do dicionário e verifica se há match com conceito do glossário


cria entrada conjunta

guarda entrada com informação proveniente de um sítio só

# Estrutura de Dados


```
{
  "conceito": "bronquiolite",
  "ca": "bronquiolitis",
  "en": [
    "bronchiolitis"
  ],
  "es": [
    "bronquiolitis"
  ],
  "eu": [
    "bronkiolitis"
  ],
  "fr": [
    "bronchiolite"
  ],
  "gl": [
    "bronquiolite",
    "bronquite capilar"
  ],
  "nl": [
    "bronchiolitis"
  ],
  "oc": [
    "bronquiolitis",
    "bronquitis capillar"
  ],
  "área médica": "Diagnòstic",
  "definicao catalã": "Inflamació dels bronquíols.",
  "significado": "Doença que se caracteriza por uma inflamação nos bronquíolos e que,
  "contexto": "...O Vírus Respiratório Sincicial provoca febre, corrimento nasal, tosse
},
```

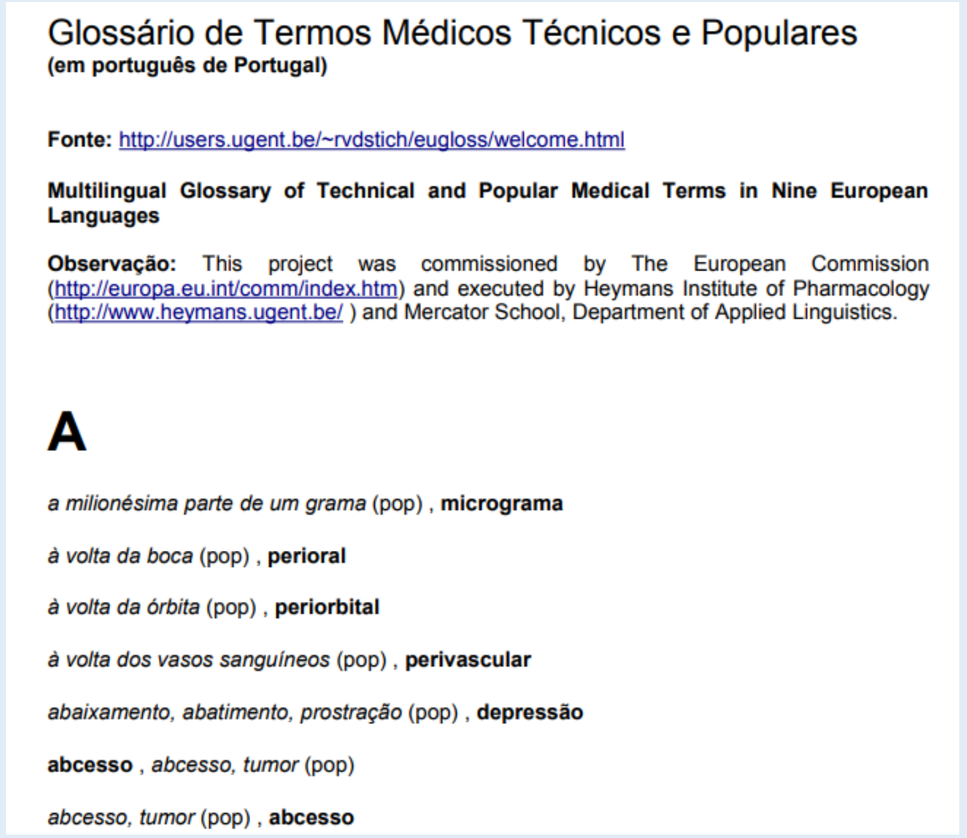
```
  "conceito": "xenotransplante",
  "en": "xenotransplantation",
  "es": "xenotransplante",
  "significado": "Tipo de transplante que envolve o uso de órg
  "contexto": "...O fato é importante porque abre a possibilidade de s
}
```

```
{
  "conceito": "antigénio",
  "sinónimos pt": [
    "antígeno"
  ],
  "ar": [
    "المستضد"
  ],
  "ca": "antigen",
  "en": [
    "antigen"
  ],
  "es": [
    "antígeno"
  ],
  "fr": [
    "antigène"
  ],
  "gl": [
    "antíxeno"
  ],
  "oc": [
    "antigèn"
  ],
  "área médica": "Epidemiologia",
  "definicao catalã": "Molècula capaç de produir una  resposta immunitària en l
}
```

# Glossário de Termos Médicos Técnicos e Populares

- Simples em termos de estrutura
- Expressões populares associadas a conceitos técnicos
- Organizado em ordem alfabética

- Entradas repetidas
- Existe ambiguídade em algumas entradas



Glossário de Termos Médicos Técnicos e Populares
**(em português de Portugal)**

**Fonte:** http://users.ugent.be/~rvdstich/eugloss/welcome.html

**Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages**

**Observação:** This project was commissioned by The European Commission (http://europa.eu.int/comm/index.htm) and executed by Heymans Institute of Pharmacology (http://www.heymans.ugent.be/ ) and Mercator School, Department of Applied Linguistics.

## A

*a milionésima parte de um grama* (pop) , **micrograma**

*à volta da boca* (pop) , **perioral**

*à volta da órbita* (pop) , **periorbital**

*à volta dos vasos sanguíneos* (pop) , **perivascular**

*abaixamento, abatimento, prostração* (pop) , **depressão**

**abcesso** , *abcesso, tumor* (pop)

*abcesso, tumor* (pop) , **abcesso**

# Tratamento do Glossário de Termos Médicos Técnicos e Populares

*pdftotxt* ⟶ Ficheiro texto ⟶ Expressões regulares :

- Retirar Form Feed
- Colocar todas as entradas seguidas
- Normalizar a divisão entre as palavras/expressões

```
texto=re.sub( pattern: r"\f", repl: "",texto) #tira as quebras de pagina
texto=re.sub( pattern: r"\n\n", repl: "\n",texto) #poe_todo o texto corrido
texto = re.sub( pattern: r"\s*[,;]\s*",  repl: ",", texto)  # ficar tudo normalizado
```

# Tratamento do Glossário de Termos Médicos Técnicos e Populares

Entradas que ocupavam mais do que uma linha não apresentavam nenhuma característica textual que as diferenciavam de uma entrada completamente distinta.

```
M
@má digestão (pop),dispepsia
@maceração,empolamento por contacto com líquidos,extracção de drogas por humedecimento,extracção a frio (pop)
@maciço,grande,amplo (pop)
```

Utilizar o índice alfabético para adicionar uma marca às entradas que de facto inicializavam pela letra do índice em que se situavam.

# Tratamento do Glossário de Termos Médicos Técnicos e Populares

Algumas entradas, que estavam formatadas de forma diferente, alteraram a ordem das palavras, separando-as também por new lines, deixando de ser percetível.

```
produção
excessiva
hiperaldosteronismo

de

aldoresterona

pela

glândula

supra-renal

(pop)

,
```

Por serem acontecimentos raros, foram tratadas manualmente

# Tratamento do Glossário de Termos Médicos Técnicos e Populares

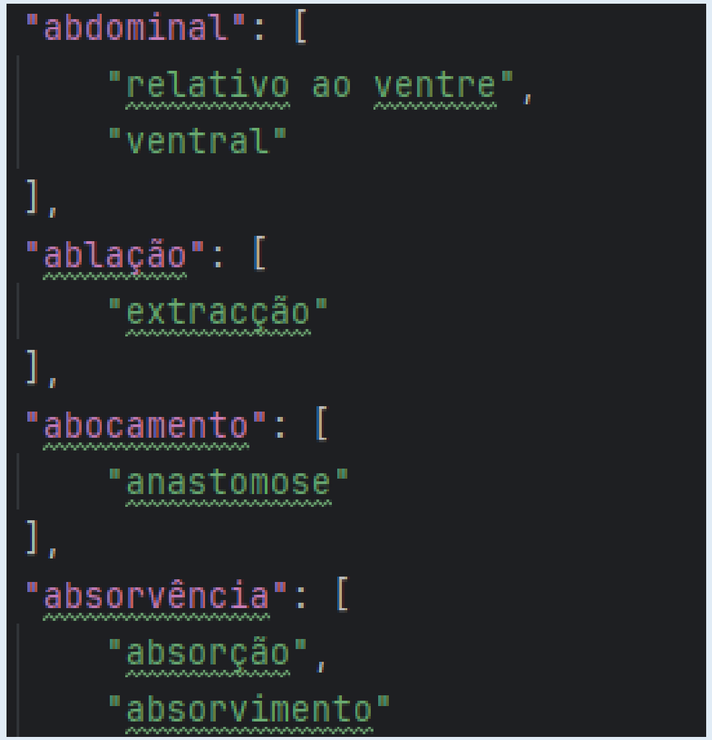
Foi criado um ficheiro json temporário para armazenar toda a informação tratada deste documento:

- Entradas repetidas : Não foram adicionadas
- Ambíguidade em entradas :
  - Entradas do tipo **ambiente** , (pop) foram removidas
  - As palavras em negrito não são esclarecedoras, logo foi feito só a divisão entre palavras/expressões populares e vocabulário técnico

# Tratamento do Glossário de Termos Médicos Técnicos e Populares

Foi criado um ficheiro json temporário para armazenar toda a informação tratada deste documento:

Dicionário
Chaves : Palavras/expressões populares
Valores : Lista de vocabulário técnico



# Tratamento do Glossário de Termos Médicos Técnicos e Populares

Foi adicionado a informação recolhida ao json final:

- Utiliza expressões regulares para ver quando alguma palavra é referenciada (ignorando palavras que não acrescentam conhecimento)

- Nesse caso, adiciona vocabulário associado

# Estrutura dos Dados

O ficheiro json com toda a informação relevante de todos os documentos, será composta por uma lista de dicionários, cada dicionário terá a seguinte estrutura:

"conceito": "neologismo",
Língua :

"*traducao_na_respeitva_Lingua_1*","*traducao_na_respeitva_Lingua_2*",...

"significado": "definição clara e contextualizada do conceito",
"contexto": "exemplo textual de uso do conceito",
"outras associacoes a 'termo referente'":

"*termo_relacionado_1*","*termo_relacionado_2*",...

# Estrutura dos Dados


```
{
  "conceito": "hantavirose",
  "en": "hantavirosis",
  "es": "hantavirosis",
  "significado": "Infecção causada no ser humano pelo hantavírus.  « a  hantavirose  é  capaz  de  matar  num  prazo
  "significado_enciclopédico": "a  hantavirose  é  capaz  de  matar  num  prazo  dois  a  três  dias  grande parte da
  "contexto": "...Minas  Gerais,  Mato  Grosso  e  Goiás  -  onde  houve  casos  de  hantavirose  em  seres  humanos
  "outras associacoes a 'infecção'": [
    "infecção"
  ],
  "outras associacoes a 'parte'": [
    "fracção",
    "porção",
    "número fraccionário (número) quebrado"
  ]
},
```

# Conclusão

- Utilizando técnicas abordadas nas aulas de PLN, utilizamos várias técnicas de modo a obter ficheiros JSON com informação proveniente de vários documentos, e acessível para manipulação futura.

# Obrigado!