

# Humanities Going Digital (HUGOD)

[2020-1-CZ01-KA226-HE-094363]

Ferramentas computacionais  
de exploração de *Corpora*:  
introdução

Sessão 02

28/09/2022

Raquel Amaro



# AULAS

Introdução



01

02

Linguística de  
*Corpus*



Ferramentas  
computacionais



03

Digitalização



04

OCR



05

Compilação de  
*Corpora*



06

## Ferramentas computacionais para exploração de *Corpora*

O que nos permitem?

- Desconstrução do texto (ou conjunto de textos) em poucos segundos → fragmentação
- Análise quantitativa dos dados (frequências de ocorrências, distribuição etc.)
- Dados estruturados e reutilizáveis para diferentes tipos de estudo → muito usados para estudos sobre a língua, mas não só.

Importante: Permite análise crítica dos dados!

- O facto de não termos evidências relativamente a um determinado fenómeno, não implica que este não exista, pode apenas não constar no corpus recolhido → não podemos inferir de forma direta “evidências negativas”

## Tratamento de corpora: ferramentas e procedimentos

- Separação de unidades lexicais: [tokenizador](#)
- Lematização verbal: [lematizador](#)
- Anotação morfossintática - [etiquetador/anotador](#)
- Análise quantitativa (frequência de ocorrência, densidade lexical, contagem de caracteres/unidades lexicais/sílabas/orações.. )
- Concordanceamento - concordanciadores para seleção de contextos de ocorrências anotados
- Extração de palavras-chave
- ...

Para que serve?

- Indicar onde começa e acaba uma palavra/unidade de significado/unidade lexical:
  - → nem sempre o que está entre dois espaços em branco é a palavra, depende da língua em uso (ex. línguas aglutinantes, línguas com ideogramas etc.)
- para identificar as palavras compostas como único itens atômicos
  - (couve-flor; rés-do-chão)

# Exemplos de Tokenizadores

6



Exemplo

Processar ficheiros

Notebook

Web Service

Documentação

Dentro deste parágrafo, há vários casos especiais para a separação de palavras na ortografia do português. Tu também deste este exemplo: dar-se-lho-ia. E prà frente é que é pelo caminho certo, etc.

Limpar

Separar

<p> <s> Dentro de\_ este parágrafo ;\*/ há vários casos especiais para a separação de palavras em\_ a ortografia de\_ o português ;\*/ </s> <s> Tu também deste este exemplo ;\*/ dar-CL-ia -se -lhe\_ -o ;\*/ </s> <s> E para\_ a frente é que é por\_ o caminho certo ;\*/ etc ;\*/ </s> </p>



<https://portulanclarin.net/workbench/lx-tokenizer/>



SegmentAnt

File Help

Input Text

Clear

语料库语言学  
维基百科,自由的百科全书  
语料库语言学是基于语言运用的实例的语言研究。语料库语言学可以对自然语言进行语法与句法分析,还可以研究它与其他语言的关系。语料库最初由手工完成,而现在主要是由电子计算机自动完成。  
语料库语言学家相信,可靠的语言分析需建立在新鲜的语料、自然的语言环境,和最小的实验干扰之上。在语料库语言学中,语料标注的意义众说纷纭,从约翰·辛克莱主张最少量的标注,并允许文本“为自己说话”,到“英语用法调查组”(设在伦敦大学学院)[2]鼓励更多的标注,并认为它是通向更完备和严谨的语言理解的道路。  
历史  
现代语料库语言学的个里程碑是亨利·库切拉和W.纳尔逊·弗朗西斯在年出版的《当代美语的句法分析》一书。该项工作基于对布朗语料库的分析。布朗语料库是一个精心编制的美国英语语料库,规模约有一百万词次。库切拉和弗朗西斯将这些语料用于各种计算分析,获得了丰富和多样化的成果,该成果结合了语言学、语言学、心理学、统计学,和社会学元素。另一关键出版物是年伦道夫·夸克的《当代英语语法》在这本书中他介绍了“英语用法调查”项目。  
此后不久,波士顿出版商霍顿·米夫林·纳许邀请库切拉为其新的美国传统英语字典提供百万词次、三级引文的来进行词典编纂。《美国传统英语字典》创新地将规定性元素(应如何使用语言)和描述性元素(语言实际上是如何被使用)结合在一起。  
其他出版社纷纷效仿,英国出版商柯林斯单语学习词典,就是为非英语母语者学习英语而出版的,它使用了“英语银行”语料库。“英语用法调查”语料库被用于由夸克等人编著的《综合英语语法》



Output Results

Clear

语料库 语言学  
维基百科 自由 的 百科全书  
语料库 语言学 是 基于 语言 运用 的 实例 的 语言 研究 语料库 语言学 可 以 对 自然语言 进行 语法 与 句法分析 还 可 以 研究 它 与 其他 语言 的 关系 语料库 最初 由 手 工 完成 而 现在 主要 是 由 电子 计算机 自动 完成  
语料库 语言学家 相信 可靠 的 语言 分析 需 建立 在 新鲜 的 语料 自然 的 语言 环境 和 最小 的 实验 干扰 之 上 在 语料库 语言学 中 语料 标注 的 意义 众说纷纭 从 约翰 辛 克莱 主张 最少 量 的 标注 并 允许 文本 为 自己 说话 到 英语用法 调查组 设 在 伦敦大学 学院 [ 2 ] 鼓励 更多 的 标注 并 认为 它是 通向 更 完备 和 严谨 的 语言 理解 的 道路  
历史  
现代 语料库 语言学 的 个 里程碑 是 亨利 库切 切拉 和 W. 纳尔 逊· 弗朗 西 斯 在 年 出版 的 当代 美语 的 句法 分析 一 书 该项 工作 基于 对 布朗 语料库 的 分析 布朗 语料库 是 一个 精心 编制 的 美国 英语 语料库 规模 约 有 一 百万 词次 库切 切拉 和 弗 朗 西 斯 将 这些 语料 用于 各种 计算 分析 获得 了 丰富 和 多样化 的 成果 该 成果 结合 了 语言学 语言学 心理学 统计学 和 社会学 元素 另一 关键 出版物 是 年 伦道夫 夸克 的 当代 英语语法 在 这 本书 中 他 介绍 了 英语 用法 调查 项目  
此后 不久 波士 顿 出版 商 霍顿 米夫 林 纳许 邀请 库切 切拉 为 其 新 的 美国 传统 英语 字典 提供 百万 词次 三级 引文 的 来 进行 词典 编纂 美国 传统 英语 字典 创 新 地 将 规定 性 元素 应 如何 使用 语言 和 描述 性 元素 语言 实际 上 是 如何 被 使用 结合 在 一 起  
其他 出版 社 纷纷 效仿 英国 出版 商 柯林 斯 单语 学习 词典 就是 为 非 英语 母语 者 学习 英语 而 出版 的 它 使用 了 英语 银行 语料库 英语用法 调查 语料库 被 用于 由 夸克 等 人 编著 的 综合 英语语法



<https://www.laurenceanthony.net/software/segmentant/>

# Lematizador

Lema = forma canônica/neutra das palavras (entrada do dicionário)

Permite:

- agrupar diferentes ocorrências de uma mesma unidade lexical
- distinguir palavras homógrafas

Para que serve?

→ para cálculo de frequência

→ para desambiguação

<https://www.linguateca.pt/acesso/corpus.php?corpus=CONDIV>

“casa” - distribuição por lemas



## Example Usage

Running the `LemmaProcessor` requires the `TokenizerProcessor`, `MWTPProcessor`, and `POSPProcessor`. After the pipeline is run, the `Document` will contain a list of `Sentence`s, and the `Sentence`s will contain lists of `Word`s. The lemma information can be found in the `lemma` field of each `Word`.

## Accessing Lemma for Word

Here is an example of lemmatizing words in a sentence and accessing their lemmas afterwards:

```
import stanza

nlp = stanza.Pipeline(lang='en', processors='tokenize,mwt,pos,lemma')
doc = nlp('Barack Obama was born in Hawaii.')
print(*[f'word: {word.text+" "}\tlemma: {word.lemma}' for sent in doc.sentences for word in sent.words], sep='')
```

As can be seen in the result, Stanza lemmatizes the word *was* as *be*.

word: Barack	lemma: Barack
word: Obama	lemma: Obama
word: was	lemma: be
word: born	lemma: bear
word: in	lemma: in
word: Hawaii	lemma: Hawaii
word: .	lemma: .



Exemplo

Documentação

ter-se-lhas-ia apresentado

Clicar para caracteres especiais:

á ã ç ê é í õ ô ú ü

Limpar

apenas verbos conhecidos

Lematizar

**teria apresentado (+ se + lhe + as)**

**apresentar**

indicativo | futuro do pretérito composto | 3ª pessoa | singular

cjg >>



# Etiquetador

Etiquetador morfossintático: ferramenta que atribui a cada palavra num texto a categoria morfossintática correspondente.

POS (tagging) = Part-of-speech - categoria morfossintática

Para que serve?

- para efeitos de classificação das unidades lexicais que compõem o corpus
- para desambiguação

<https://www.linguateca.pt/acesso/corpus.php?corpus=CONDIV>

“clara” distribuição por POS

“corte” distribuição por POS

# Exemplo de Etiquetador

10



Time taken: 0.8848 mc

```
Arte_PROPN Pública_PROPN . PUNCT Origens_NOUN e_CCONJ condição_NOUN histórica_ADJ , PUNCT José_PROPN Guilherme_PROPN Abreu_PROPN . PUNCT
1_NUM . PUNCT Introdução_NOUN
O_DET tema_NOUN da ADP origem_NOUN da ADP arte_NOUN pública_ADJ não_ADV tem AUX sido_AUX um_DET objeto_NOUN de ADP estudo_NOUN da ADP histo.
Assim_ADV sendo AUX , PUNCT importa_VERB desde ADP já_ADV referir_VERB que_SCONJ apesar_ADV do_DET seu_DET titulo_NOUN algo_ADV ambicioso_AI
2_NUM . PUNCT Excerto_NOUN etimológico_ADJ . PUNCT
Antes_ADV de ADP entrar_VERB no ADP estudo_NOUN histórico_ADJ propriamente_ADV dito_VERB , PUNCT importa_VERB interrogar_VERB a_DET origem_I
Etimologicamente_ADV , PUNCT público_NOUN é AUX uma_DET palavra_NOUN que PRON deriva_NOUN do ADP latim_NOUN clássico_ADJ " PUNCT publicus_NOUN
O_DET processo_NOUN de ADP formação_NOUN do ADP vocábulo_NOUN publicus_NOUN decorre_VERB , PUNCT por_NOUN sua ADP vez_NOUN , PUNCT da ADP a
. PUNCT Enquanto_ADV coisa_NOUN da ADP população_NOUN designa_VERB o PRON que PRON é AUX comum ADJ e_CCONJ partilhável_ADJ por ADP todos_PRON
. PUNCT Enquanto_ADV coisa_NOUN do ADP povo_NOUN designa_VERB o PRON que PRON é AUX comum ADJ e_CCONJ partilhado_VERB coletivamente_ADV . P
. PUNCT Enquanto_ADV coisa_NOUN dirigida_VERB à ADP população_NOUN designa_VERB uma_DET comunicação_NOUN destinada_VERB a ADP todos_PRON .
. PUNCT Enquanto_ADV coisa_NOUN oposta_ADJ ao ADP privado_ADJ designa_VERB oposição_NOUN à ADP exclusão_NOUN da ADP partilha_NOUN do PRON q

Função_NOUN mediadora_ADJ da ADP Esfera_PROPN Cívica_PROPN

Sentido_NOUN de ADP identidade_NOUN . PUNCT
O_DET sentido_NOUN polissêmico_ADJ da ADP noção_NOUN de ADP público_NOUN reflete-se_VERB na ADP linguística_NOUN de ADP forma_NOUN muito_AD
A- ADP Como ADP substantivo_ADJ , PUNCT público_NOUN designa_VERB : PUNCT
. PUNCT 1_NUM . PUNCT A_DET comunidade_NOUN ou_CCONJ povo_NOUN como ADP uma_DET totalidade_NOUN . PUNCT Ex_NOUN . PUNCT : PUNCT os_DET cida
. PUNCT 2_NUM . PUNCT Um_DET setor_NOUN da ADP população_NOUN que PRON compartilha_VERB um_DET interesse_NOUN comum ADJ . PUNCT Ex_NOUN . P
. PUNCT 3_NUM . PUNCT Os_DET admiradores_NOUN ou_CCONJ seguidores_NOUN de ADP uma_DET pessoa_NOUN famosa_ADJ . PUNCT Ex_NOUN . PUNCT : PUNCT
B- PROPN Como ADP adjetivo_NOUN , PUNCT público_NOUN discrimina_VERB : PUNCT
. PUNCT 1_NUM . PUNCT _SPACE O PRON que PRON pertence_VERB , PUNCT afeta_VERB ou_CCONJ se PRON relaciona_VERB com ADP a_DET comunidade_NOUN
. PUNCT 2_NUM . PUNCT _SPACE O PRON que PRON é AUX mantido_VERB e/ NOUN ou_CCONJ usado_VERB pelo ADP povo_NOUN ou_CCONJ pela ADP comunida
. PUNCT 3_NUM . PUNCT _SPACE O PRON que PRON é participado_VERB ao ADP povo_NOUN ou_CCONJ à ADP comunidade_NOUN . P
. PUNCT 4_NUM . PUNCT _SPACE O PRON que PRON se PRON fag_VERB em ADP benefício_NOUN do ADP povo_NOUN , PUNCT da ADP comunidade_NOUN ou_CCO
. PUNCT 5_NUM . PUNCT _SPACE O PRON que PRON se PRON integra_VERB ou_CCONJ funciona_VERB numa ADP escola_NOUN pública_ADJ . PUNCT Ex_NOUN
. PUNCT 6_NUM . PUNCT _SPACE O PRON que PRON é AUX do ADP conhecimento_NOUN e_CCONJ juízo_PROPN de ADP todos_PRON . PUNCT Ex_INTJ . PUNCT
C- PROPN Como ADP verbo_NOUN , PUNCT publicar_VERB significa_VERB : PUNCT
. PUNCT 1_NUM . PUNCT Tornarpatente_ADV e_CCONJ manifestar_VERB algo PRON ao ADP público_NOUN . PUNCT Ex_INTJ . PUNCT : PUNCT a_DET publici
. PUNCT 2_NUM . PUNCT Revelar_VERB ou_CCONJ digir_VERB o PRON que PRON estava AUX secreto_ADJ ou_CCONJ oculto_ADJ . PUNCT Ex_INTJ . PUNCT :
. PUNCT 3_NUM . PUNCT Difundirpor_PROPN meio_PROPN da ADP imprensa_NOUN ou_CCONJ de ADP qualquer_DET outro_DET procedimento_NOUN técnico_AI

3_NUM . PUNCT O_DET Conceito_PROPN de ADP Público_PROPN em ADP Hannah_PROPN Arendt_PROPN
Sobre ADP conceito_NOUN de ADP público_NOUN , PUNCT a_DET definição_NOUN de ADP Hannah_PROPN Arendt_PROPN expressa_VERB com ADP clareza_NOUN
O_DET termo_NOUN público_ADJ denota_VERB dois NUM fenómenos_NOUN intimamente_ADV relacionados_VERB , PUNCT mas_CCONJ não_ADV completamente_I
E_CCONJ mais_ADV adiante_ADV , PUNCT acrescenta_VERB : PUNCT
Em ADP segundo_ADJ lugar_NOUN , PUNCT o_DET termo_NOUN público_ADJ significa_VERB o_DET próprio_DET mundo_NOUN , PUNCT na ADP medida_NOUN em
Em ADP Arendt_PROPN , PUNCT a_DET dialética_NOUN da ADP noção_NOUN de ADP público_NOUN opõe_VERB a_DET delimitação_NOUN passiva_ADJ do ADP
Ontologia_NOUN e_CCONJ Genealogia_PROPN de ADP Público_PROPN
Sentido_NOUN abstrato_ADJ - PUNCT Integração_NOUN - PUNCT O_DET substantivo_NOUN
Público_PROPN é AUX o PRON que PRON é AUX comum ADJ ( PUNCT o_DET ser_NOUN : PUNCT o_DET conceito_NOUN não_ADV inscrito_VERB ) PUNCT
Sentido_NOUN de ADP atividade_NOUN - PUNCT síntese_NOUN operativa_ADJ . PUNCT o_DET verbo_NOUN
Publicar_VERB é AUX comunicar_VERB o PRON que PRON é PUNCT comum ADJ ( PUNCT o_DET modo_NOUN de ADP ser_VERB : PUNCT o_DET inscrever_VERB )
Sentido_NOUN concreto_ADJ - PUNCT diferenciação_NOUN - PUNCT o_DET adjetivo_NOUN
Público_PROPN é AUX o PRON que PRON é AUX partilhado_ADJ ( PUNCT o_DET ente_NOUN : PUNCT o_DET conceito_NOUN inscrito_VERB ) PUNCT
4_NUM . PUNCT Complexo_NOUN conceptual_ADJ de ADP Arte_PROPN Pública_PROPN
```



<https://www.laurenceanthony.net/software/tagant/>

# Exemplo de Etiquetador

11



Exemplo

Processar ficheiros

Notebook

Web Service

Documentação

Etiquetas

Arte Pública. Origens e condição histórica, José Guilherme Abreu.

O tema da origem da arte pública não tem sido um objeto de estudo da história da arte, e se nos últimos anos o estudo do mesmo tem conhecido um desenvolvimento importante na bibliografia de língua inglesa, assim como em castelhano e mesmo em português, a investigação tem-se centrado essencialmente sobre casos de estudo, desde obras, projetos ou intervenções autorais, estendendo-se mais raramente a programas de regeneração urbana ou de participação comunitária, onde são analisadas sobretudo as linguagens plásticas, as estratégias de produção artística e as tensões causadas pela sua receção pública das obras, sendo residuais os trabalhos sobre os problemas e os conceitos de uma teoria da arte pública, que globalmente está por estabelecer.

Formato de visualização  amigável  CINTIL  CONLL  coluna

Limpar

Anotar

<p> <s> PNM PNM PNT . </s> <s> CN CJ CN ADJ PNT PNM PNM PNM PNT . </s> <s> DA CN PREP  
Arte Pública . Origens e condição histórica , José Guilherme Abreu . O tema de\_  
DA CN PREP DA CN ADJ ADV VAUX PPT UM CN PREP CN PREP DA CN PREP DA CN PNT CJ CL PREP  
a origem de\_ a arte pública não tem sido um objeto de estudo de\_ a história de\_ a arte , e se em\_  
DA ADJ ANS DA CN PREP DA ADV VAUX PPT UM CN ADJ PREP DA CN PREP  
os últimos anos o estudo de\_ o mesmo tem conhecido um desenvolvimento importante em\_ a bibliografia de  
CN ADJ PNT LCJ1 LCJ2 PREP CN CJ ADV PREP CN PNT DA CN V CL PPT  
língua inglesa , assim como em castelhano e mesmo em português , a investigação tem -se centrado  
ADV PREP CN PREP CN PNT PREP CN PNT CN CJ CN ADJ PNT GER CL ADV  
essencialmente sobre casos de estudo , desde obras , projetos ou intervenções autorais , estendendo -se mais  
ADV DA CN PREP CN ADJ CJ PREP CN ADJ PNT REL V PPA ADV  
raramente a programas de regeneração urbana ou de participação comunitária , onde são analisadas sobretudo  
DA CN ADJ PNT DA CN PREP CN ADJ CJ DA CN PPA PREP DA POSS CN ADJ  
as linguagens plásticas , as estratégias de produção artística e as tensões causadas por\_ a sua receção pública  
PREP DA CN PNT GER ADJ DA CN PREP DA CN CJ DA CN PREP UM CN PREP DA CN  
de\_ as obras , sendo residuais os trabalhos sobre os problemas e os conceitos de uma teoria de\_ a arte  
ADJ PNT REL ADV V PREP INF PNT  
pública , que globalmente está por estabelecer . </s> </p>



<https://portulanclarin.net/workbench/lx-tagger/>

# Análise quantitativa

Análise objetiva de dados (totais ou parciais) ou parâmetros específicos

Ex.

- lista de ocorrências (=wordlist)
- nº de frases, parágrafos, lemas, formas etc..
- tokens por texto
- distribuição do lema ou da expressão pelo corpus (com diferentes visualizações: diagramas, plots etc.)

# Exemplo de Análise Quantitativa

13



Exemplo Processar ficheiros Notebook Web Service Documentação

A final do Campeonato Europeu de Futebol de 2016 realizou-se em 10 de julho de 2016 no Stade de France em Saint-Denis, França. Foi disputada entre Portugal e a França, que era a equipa anfitriã. Os portugueses ganharam a partida e sagraram-se campeões europeus de futebol. Esta foi a segunda participação numa final deste campeonato para Portugal e a terceira para a França. Os portugueses haviam participado anteriormente nas edições de 1984 e em todas as edições desde 1996. O seu melhor resultado anterior foi em 2004, com o título de vice-campeão. Já os franceses participaram em 1960, 1984 e em todas as edições desde 1992, tendo-se sagrado campeões nas edições de 1984 e de 2000.

Limpar Analisar

Ocorrências de letras: 559  
Média de letras por palavra: 4,47

Número de sílabas: 243  
Média de sílabas por palavra: 1,94

Ocorrências de palavras: 125  
Média de palavras por frase: 17,86  
Proporção de palavras únicas: 52,00%

Frases: 7  
Orações simples: 11  
Orações passivas: 0 (0,00%)  
Orações subordinadas: 1 (9,09%)  
Orações coordenadas: 6

Índice de Flesch: 24,25

Densidade lexical:

Verbos	12	8,76%
Nomes	20	14,60%
Adjetivos	4	2,92%
Advérbios	2	1,46%
Preposições	28	20,44%

Frequências de palavras:

de	12
em	10
a	7
e	6
as	4
edições	4
o	4
1984	3
foi	3
frança	3
os	3
se	3
2016	2
campeonato	2
campeões	2
desde	2
final	2
futebol	2
para	2
portugal	2
portugueses	2
todas	2
10	1
1960	1
1992	1
1996	1
2000	1
2004	1
anfitriã	1
anterior	1
anteriormente	1
com	1
disputada	1



<https://portulanclarin.net/workbench/lx-quantitative/>

# Concordanceador

- Concordanceador: ferramenta que permite fazer pesquisas dentro do corpus e identificar e recuperar uma sequência específica de caracteres mais ou menos comprida - uma palavra, uma parte de palavra, um sintagma etc.
- Concordância: linha de texto em que ocorre a sequência de caracteres em causa. Tipicamente para cada ocorrência corresponde uma linha.

## Que tipo de resultado fornece?

A sequência em causa está sempre evidenciada (negrito, cor diferente), normalmente é apresentada no meio da linha, mas não necessariamente. O contexto antes e depois da forma em questão, pode corresponder com uma frase completa ou com parte desta.

## Para que serve?

- para analisar o contexto de ocorrência de formas ou sequências (é possível aplicar filtros para isolar co-ocorrentes mais ou menos distantes antes e depois da forma em causa)
- para identificar padrões linguísticos recorrentes
- para inferir informações - confirmar ou negar hipóteses prévias

Procura: [lema="jogador"]  
Pedido de uma concordância em contexto  
Corpo: CONDIVport 11.2

8090 ocorrências.

Número de ocorrências excessivo! Tente restringir a sua procura a menos de 4000 casos.

### Concordância

Procura: [lema="jogador"].

Apresenta-se uma amostra aleatória de 4000 das 8090 ocorrências encontradas.

*par=fut-PT-Bola-50-5912*: Dos três defesas, Alfredo continua a ser o mais **jogador** -- um jogador autêntico, aliás .

*par=fut-PT-Bola-50-6215*: Os recortes de alguns jornais que damos, a seguir devem habilitar os nossos leitores, a ajuizar melhor do reflexo causado pela decisão tomada por Cândido de Oliveira, e da forma como foram recebidas pelos **jogadores**, críticos e dirigentes as suas primeiras intervenções como orientador técnico do importante clube carioca .

*par=fut-BR-GazetaEsp-50-60003*: Até mesmo antes da leitura dos jornais, fervilhavam os dos «fans» em torno da atuação dos **jogadores**, cada qual, como é natural, procurando colocar em o seu idolo .

*par=fut-BR-GazetaEsp-70-63841*: — Ainda não sei se poderei contar com o **jogador** .

*par=fut-BR-GazetaEsp-70-62295*: **Jogadores** -- Armando, Osvaldo Cunha, Mendes, Vagner, Pedro Rodrigues, Ademir, Adãozinho, Peri, Servilio, Rocha, Lima, Mauri, Leonete, Ico, Nelson Lopes, José Eduardo, Neilo e Carlos Alberto .

*par=fut-PT-Bola-50-2563*: -- Em quatro meses de permanente labor, não registámos a mais pequena deselegância da parte dos **jogadores** .

*par=fut-PT-Bola-50-9653*: **Jogadores** correctos e público correctíssimo valorizaram o espectáculo o mais possível .

*par=fut-PT-Bola-50-8988*: Aproveitámos a ocasião para ouvir alguns **jogadores** que se dirigiam para o balneário .

*par=fut-PT-Bola-70-17741*: Coexistirão na equipa jogadores não amadores e **jogadores** profissionais ?

*par=fut-BR-JSports-50-37601*: Quando joga o scratch inglês os **jogadores** do scratch abandonam os seus clubes até em dias de matches decisivos do campeonato inglês .

*par=fut-BR-JSports-70-43169*: Mas, ainda que já tenha o seu time na cabeça, Zagalo enfrenta dois problemas sérios, resultantes de um mesmo **jogador**: Tostão .

*par=fut-PT-Bola-50-6102*: Pelas substituições operadas ao longo da partida, verificou-se que passaram pelo campo 33 **jogadores**, dos quais dezanove pertencentes ao F. C do Porto e catorze ao Estoril .

*par=fut-PT-Bola-50-582*: -- A sua transferência para uruchiba da III Divisão «afogou-o»... mas ele sabe «nadar» muito bem O contacto com **jogadores** das Divisões inferiores fez-lhe perder qualidades, mas não popularidade...

*par=fut-PT-Bola-70-13291*: Finalidade para a qual, naturalmente, terá também contribuído o árbitro, ao ordenar a marcação de duas grandes penalidades, ambas por derrube de **jogadores** vimaranenses .

*par=fut-BR-JSports-70-41332*: Em seu trajeto, os **jogadores** eram aplaudidos .



## BNCweb (CQP-Edition)

Your query "time" returned 152502 hits in 3860 different texts (98,313,429 words [4,048 texts]; frequency: 1551.18 instances per million words) (0.289 seconds)

No	Filename	Hits 1 to 50 Page 1 / 3051
1	<a href="#">A00_81</a>	Many people with AIDS have to spend long periods of <b>time</b> in hospital unless there is someone at home who can help and look after them.
2	<a href="#">A00_82</a>	ACET volunteers work as part of a team and provide help in many different ways to ensure that people don't spend <b>time</b> in hospital unnecessarily.
3	<a href="#">A00_88</a>	How much <b>time</b> to I need to give?
4	<a href="#">A00_134</a>	Dr Dixon said, 'With up to 20 years from infection to illness, we just have to ask how many of our congregation have been added during that <b>time</b> ?'
5	<a href="#">A00_158</a>	It was the first <b>time</b> our national and international network had gathered together in one place and made us all realise just how much the work has grown.'
6	<a href="#">A00_206</a>	By working co-operatively, long-term, with the people around me, I hope to continue for some <b>time</b> yet.
7	<a href="#">A00_223</a>	After spending some <b>time</b> , often years, with us, people move on to other things and we need to fill the gaps.'
8	<a href="#">A00_226</a>	In London, volunteer training programmes will now take place every September and February and will require an individual's <b>time</b> for one evening a week over a six-week programme.
9	<a href="#">A00_252</a>	The biggest changes are in the length of <b>time</b> people ill with the disease are now surviving and in the nature of the illnesses themselves.
10	<a href="#">A00_275</a>	These factors help explain some of the reasons why the total number of ACET clients covered at any one <b>time</b> by our on call service in London has more than doubled from 70 in April 1990 to over 150 by March 1991; and why the nature of the services required has become so much more sophisticated.
11	<a href="#">A00_340</a>	By working co-operatively, longterm, with the people around me, I hope to continue for some <b>time</b> yet.
12	<a href="#">A00_363</a>	Within a short space of <b>time</b> referrals were regularly coming in.
13	<a href="#">A00_392</a>	He is not at all well, very breathless, and by the <b>time</b> we are in the care he is gasping for air.
14	<a href="#">A00_401</a>	4.00pm — We set off again; this <b>time</b> via Tony's home to collect a variety of possessions, finally arriving at hospital no.3.
15	<a href="#">A01_61</a>	Over a period of <b>time</b> they will all be ill.
16	<a href="#">A01_146</a>	Strong friendship takes <b>time</b> to build.
17	<a href="#">A01_148</a>	Friendship takes <b>time</b> .
18	<a href="#">A01_149</a>	In most happy marriages, husband and wife continue to make <b>time</b> to be with each other, and to understand each other .
19	<a href="#">A01_188</a>	Without ACET's practical support at home they could spend long periods of <b>time</b> in hospital unnecessarily.
20	<a href="#">A01_189</a>	The reality of AIDS is that the person can die at any <b>time</b> .
21	<a href="#">A01_247</a>	At the <b>time</b> you enter a Deed of Covenant, the covenant should be capable of lasting for more than 3 years , and there should be the intention by you that it does so.
22	<a href="#">A01_256</a>	If you use the covenant form attached to this leaflet you can make each annual payment at any <b>time</b> , or by any instalments, you wish as long as you make the full annual payment by the end of each 12-month period.
23	<a href="#">A01_340</a>	The intention of a Gift Aid scheme is to encourage giving without the donor being tied to a particular charity for this length of <b>time</b> .
24	<a href="#">A01_342</a>	Under Gift Aid there are no formalities at the <b>time</b> of the gift, just a cheque or cash gift to the charity.
25	<a href="#">A01_343</a>	The certificate on form R190(SD) must be completed but this can be done at any <b>time</b> and amounts to little more than a claim procedure.
26	<a href="#">A01_373</a>	The donor must be resident in the United Kingdom at the <b>time</b> that the gift is made.
27	<a href="#">A01_381</a>	If you wish to give small amounts regularly, e.g. monthly, you could accumulate the money in a separate account and then convert this to a gift to ACET every <b>time</b> the amount reaches £600.
28	<a href="#">A01_410</a>	Moreover, death is a <b>time</b> of great stress to those you love most.
29	<a href="#">A01_470</a>	This means you can get professional help any <b>time</b> of the day or night, and at weekends.
30	<a href="#">A01_529</a>	A recent survey of church youth groups shows that 1 in 4 have had sex by the <b>time</b> they are eighteen years old, 1 in 10 under the age of sixteen.





# Extração de palavras-chave

Processo automático utilizado para a extração de informação relevante a partir de um ou mais textos (corpora)

Tipicamente baseados em cálculos estatísticos para identificar as palavras mais relevantes dentro do input textual considerado

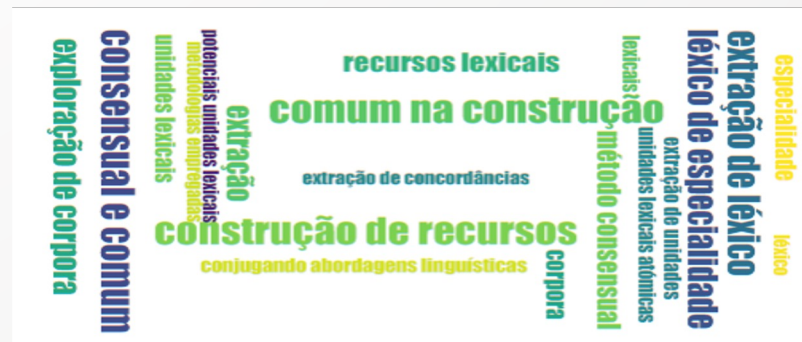
Existem 2 abordagens para extrair palavras-chaves e para treinar as ferramentas automáticas:

- por comparação com um standard (reference corpus ou golden standard), aplicando ou não algoritmos de machine-learning para melhorar constantemente o desempenho da ferramenta -> language-dependent, parcialmente domain-dependent e precisa do auxílio de ferramentas complementares (vocabulários controlados, recursos lexicais etc.) (<https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>)
- baseada em medidas estatísticas (de co-ocorrência dos elementos do próprio texto) e nas características textuais do texto ou do conjunto de textos em análise (<https://www.sciencedirect.com/science/article/pii/S0020025519308588>)

## Texto anotado

A **exploração de corpora** para a **extração de léxico** de **especialidade** é um método **consensual e comum** na **construção de recursos lexicais**. No entanto, as **metodologias empregadas** não são explicitamente discutidas, dificultando a comparação e a determinação de abordagens robustas. Para preencher esta lacuna, neste artigo apresentamos e discutimos uma metodologia detalhada para **extração de léxico** de **especialidade** a partir de **corpora**, **conjugando abordagens linguísticas** e estatísticas. O método proposto prevê tanto o uso de **corpora** de **especialidade** como de **corpora** monitores e inclui: i) análise de dados de frequência; ii) **extração** de concordâncias e colocações; iii) **extração** de informação de ordem textual, permitindo a **extração** de **unidades lexicais** atômicas e multipalavra e de relações semânticas relevantes. Como princípio base de garantia de qualidade, a proposta inclui validação dos dados finais por parte de especialistas. Deste modo, o objetivo da metodologia é a determinação de listas de potenciais **unidades lexicais** de **especialidade** e de informações relevantes para a sua descrição que permitam uma validação final rápida e eficiente, maximizando o valor informacional da interação com os especialistas.

## Nuvem de palavras



## Palavras-chave

Pontuação	n-grama
0,021910596993882824	consensual e comum
0,021910596993882824	comum na construção
0,021910596993882824	construção de recursos
0,024277956086531063	extração de léxico
0,028746159502555813	léxico de especialidade
0,03643768539079015	exploração de corpora
0,04035126286249145	extração
0,04601595761636467	método consensual
0,046663892103767375	recursos lexicais
0,05921738501646989	especialidade
0,06002427740848347	corpora
0,06835989446700896	unidades lexicais
0,0866008319943846	extração de unidades
0,09639310023377093	unidades lexicais atômicas
0,09671816241649626	conjugando abordagens linguísticas
0,10149199943030174	lexicais
0,11141158535628319	potenciais unidades lexicais
0,11476075390596947	léxico
0,1149336208382416	metodologias empregadas
0,1160654588669079	extração de concordâncias



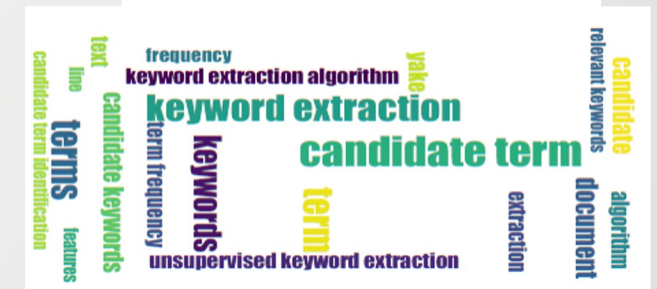
## Texto anotado

future work **Keyword extraction** algorithms have become the key components in many computer science applications. In this article, we have presented **YAKE !**, a novel feature-based online system for multi-lingual **keyword extraction** from individual documents that can be used both as a stand-alone tool and as a support for several applications, including summarization, clustering, indexing, and information visualization, to name just a few. **YAKE !** follows an unsupervised approach. It does not require a corpus or a dictionary (corpus independent), which means it can be quickly adapted across different languages or domains (domain and language independent). Our solution scales to any **document** length in a linear manner in the number of **candidate terms** identified and is **term** frequency-free, meaning that no conditions are set with respect to the minimum **frequency** or sentence **frequency** that a **candidate** keyword must have. Our experiments confirm that **YAKE !** can lead to superior **keyword extraction** results compared with ten popular state-of-the-art unsupervised **keyword extraction** algorithms, under a large number of **text** documents belonging to twenty different datasets. It also offers similar or superior effectiveness to state-of-the-art supervised methods, without the need for an extensive training set for multiple languages and domains. As a further contribution, we provide an app on Google Play, a demo, an API, and a python package so that **YAKE !** can be tested and used by the research community. **YAKE !** has already been used for temporal clustering and ranking, and won first prize at the Arquivo.pt 2018 Portuguese internet archive contest with "Tell me Stories" [archive.tellmestories.pt], a news summarization platform which generates a timeline summary for a given news topic in a storytelling-style fashion. In the future, we plan to develop a supervised solution, which will be built using the statistical **features** introduced in this work. An analysis of the importance of each feature has been already conducted in this research. Furthermore, we are currently working on an extended version of **YAKE !** to tackle the problem of absent **keywords**, for instance when a large number of manually-assigned **keywords** cannot be found in the **text** itself. We plan to do this based on word embeddings.

## Palavras-chave

Pontuação	n-grama
0,0026072961246104177	candidate term
0,003392069824422952	term
0,003422799305155864	keyword extraction
0,0036689734835595195	terms
0,004687604124718088	keywords
0,006948990799399872	document
0,007403780495156102	candidate
0,008610707179194595	candidate keywords
0,00974898610265248	yake
0,010598117626789114	text
0,010984486480505979	term frequency
0,0120421914184087	algorithm
0,012438828392054232	keyword extraction algorithm
0,012615110881503395	extraction
0,012760619978787826	unsupervised keyword extraction
0,015886390222924134	frequency
0,017680157572175777	candidate term identification
0,01787379293168973	relevant keywords
0,020651804058790486	line
0,02120125058770971	features

## Nuvem de palavras



# Corpus manager, corpus management systems ou corpus query systems

Existem ferramentas/interfaces (online ou não) que permitem explorar as potencialidades dos *corpora*, enquanto coletâneas de pistas linguísticas para análises/inferências linguísticas

Exemplos de programas a explorar:



❑ Sketch Engine (<https://www.sketchengine.eu/>)

❑ AntConc (<https://www.laurenceanthony.net/software/antconc/>)



❑ Linguatca (<https://www.linguatca.pt/>)

- ❓ Pensar num tópico de trabalho
- ❓ Pesquisar repositórios onde haja os textos relativos ao tópico
- ❓ Fazer um resumo sobre o trabalho a apresentar
- ❓ Inserir a informação relativa ao trabalho no formulário (para avaliação)

- ❑ Pensar num tópico de trabalho □ ex. vinho
- ❑ Pesquisar repositórios onde haja os textos relativos ao tópico □ revistas, sites, ...
- ❑ Fazer um resumo sobre o trabalho a apresentar □ *Observando o panorama português, pouco tem sido feito no sentido de compreender e desenvolver estudos no domínio da Análise Sensorial Enológica ou prova de vinho ou prova organoléptica. Neste sentido, o objetivo deste trabalho é compreender quais os descritores utilizados nas notas de prova provenientes das provas de vinho, que constituem um tipo de discurso especializado em Enologia, com características próprias e realizadas por especialistas.*
- ❑ Inserir a informação relativa ao trabalho no formulário (para avaliação)

# BIBLIOGRAFIA

**FALTA**

## Livros e artigos

? Xxx

? xxx

## Palestras online



? LINK

? LINK





### Trabalho de casa

- Responder a umas perguntas no questionário
- Leitura recomendada

## Avaliação

**FALTA**



### Para a próxima aula

- Pensar num tópico para desenvolver
- Escolher e recolher material para digitalizar em aula

FIM DA SESSÃO - CONTACTO PARA DÚVIDAS + HORÁRIO DE DÚVIDAS OU DE APOIO SÍNCRONO





<b>EN</b>	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
<b>PT</b>	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.