Workshop:

How to query corpora

Idalete Dias | Ana Correia | Ana Oliveira | Sílvia Araújo

Institute of Arts and Human Sciences

University of Minho

Summer School Per-Fide 2013| Corpus Linguistics and Natural Language Processing
9 - 12 September, University of Minho (Braga, Portugal)

1) Query

2) Annotation

3) Corpora

3.1.) Corpus search tools

4) Regular Expressions

5) Exercises

Query

start of a linguistic research project

verify, identify, classify examples in a corpus in order to start to develop a hypothesis or a research methodology

essential for checking results derived by automatic procedures and to examine examples in a text in more detail

Regular expression

(basic toolkit which the linguist uses)

Per-fide

Annotation

- titles, paragraphs, chapters, etc (alignment)

- metadata
- morphosyntactic tagging
- lemmatization
- prosodic annotation

...

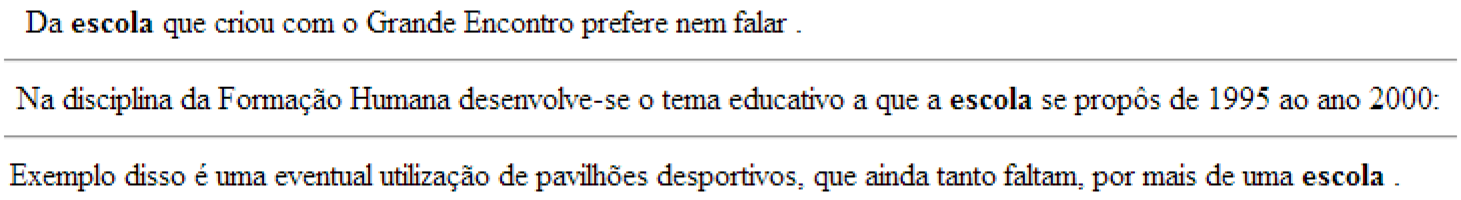
Linguistic research

optimization of search results

COMPARA http://www.linguateca.pt/COMPARA/	**OPUS** http://opus.lingfil.uu.se/bin/opuscqp.pl
Parallel (pt<->en)	Parallel (multililngual)
3 000 000 words	hundreds of millions and counting...
Literary	Technical (health, informatics, legislation, subtitles, etc.)

Corpus search tools

Concordance: provides context for your search term

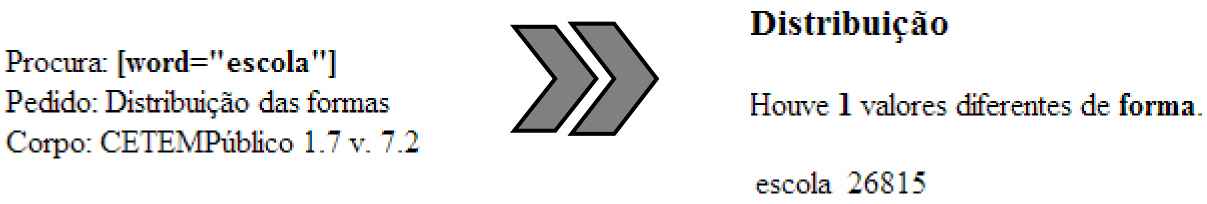


Da **escola** que criou com o Grande Encontro prefere nem falar .

Na disciplina da Formação Humana desenvolve-se o tema educativo a que a **escola** se propôs de 1995 ao ano 2000:

Exemplo disso é uma eventual utilização de pavilhões desportivos, que ainda tanto faltam, por mais de uma **escola** .

Distribution: displays frequency information



Procura: **[word="escola"]**
Pedido: Distribuição das formas
Corpo: CETEMPúblico 1.7 v. 7.2

Distribuição

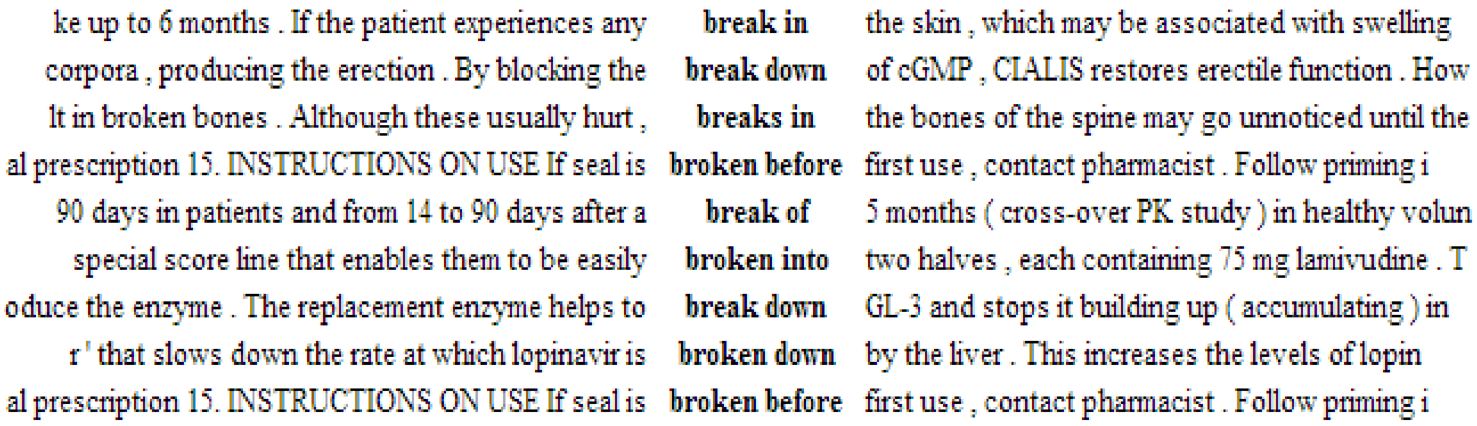
Houve **1** valores diferentes de **forma**.

escola 26815

REGULAR EXPRESSIONS (regex)

- search expression
- sequence of characters that form a **search pattern** that matches a **target string**

A regular expression like [lem="break"][pos="IN"] would match the highlighted text]



ke up to 6 months . If the patient experiences any	**break in**	the skin , which may be associated with swelling
corpora , producing the erection . By blocking the	**break down**	of cGMP , CIALIS restores erectile function . How
lt in broken bones . Although these usually hurt ,	**breaks in**	the bones of the spine may go unnoticed until the
al prescription 15. INSTRUCTIONS ON USE If seal is	**broken before**	first use , contact pharmacist . Follow priming i
90 days in patients and from 14 to 90 days after a	**break of**	5 months (cross-over PK study) in healthy volun
special score line that enables them to be easily	**broken into**	two halves , each containing 75 mg lamivudine . T
oduce the enzyme . The replacement enzyme helps to	**break down**	GL-3 and stops it building up (accumulating) in
r ' that slows down the rate at which lopinavir is	**broken down**	by the liver . This increases the levels of lopin
al prescription 15. INSTRUCTIONS ON USE If seal is	**broken before**	first use , contact pharmacist . Follow priming i

REGEX SYNTAX

- **Metacharacters**
  - special characters that have a functional value;
  - usual metacharacters are: . | * + ? [] {} ()
- **Literal characters**
  - a literal is a string we’re looking for

METACHARACTERS

- **Period / dot (.)**
  - matches any single character

Query string: [word=".do"]

looks for any 3 letter word ending with “**do**”

- “Without further **ado**, I shall give you the floor”
- “As they say, Mr President, there has been much **ado** about nothing”

(OPUS - EuroParl3)

METACHARACTERS

- **Boolean ‘or’**
  - vertical bar that separates alternatives (|)
  - alternative patterns are evaluated from left to right

Query string: [word="col(o|u)r"]

looks for an instance of “**color**”. If no instance is found, “**colour**” is searched for instead.

METACHARACTERS

- **Quantification / Iteration**
  - a quantifier after a character / group of characters specifies how often that **preceding element** is allowed to occur
  - Most common quantifiers:

    - question mark (**?**)

    - asterisk (*****)

    - plus sign (**+**)

METACHARACTERS

❑ **Quantification / Iteration**

❑ **?** = indicates there is 0 or 1 of the preceding element

Query string: [word="behaviou?r"]

matches both “**behavior**” and “**behaviour**”

METACHARACTERS

- ❑ **Quantification / Iteration**
    - ❑ ***** = indicates there is 0 or more of the preceding element

Query string: [word="oxy.*"]

finds words beginning with “**oxy-**”

➢ **oxy**gen; **oxy**genation; **oxy**clozanide; **oxy**morphone,...

(OPUS - European Medicines Agency Documents)

METACHARACTERS

- ❑ **Quantification / Iteration**
    - ❑ ***** = indicates there is 0 or more of the preceding element

Query string: [word=".*oxy"]

finds words ending with “**oxy-**”

➢ pr**oxy**; monometh**oxy**; aminoeth**oxy**; hydr**oxy**; carb**oxy**;...

(OPUS - European Medicines Agency Documents)

METACHARACTERS

- ❑ **Quantification / Iteration**
  - ❑ ***** = indicates there is 0 or more of the preceding element

Query string: [word=".*oxy.*"]

finds words with “**oxy-**” in the middle

➢ hydr**oxy**phosphate; carb**oxy**lic; cyclo-**oxy**genase;...

(OPUS - European Medicines Agency Documents)

METACHARACTERS

- ❑ **Quantification / Iteration**
  - ❑ **+** = indicates there is 1 or more of the preceding element

Query string: [word="(ha)+"]

matches “**ha**”, “**haha**”, “**hahaha**”,...

METACHARACTERS

- ❑ **Backslash (\)**
  - ❑ used to indicate that we want to use a metacharacter as a literal character
  - ❑ place the backslash in front of the metacharacter we want

Query string: [word="Dr\."]

- **Dr.**

METACHARACTERS

- **Exclamation mark (!)**
  - used to exclude elements
  - Exclamation mark preceding the equals sign means **does not equal**

‘dream’ followed by anything other than ‘about’

Query string: [lema="dream"][word!="about"]

- “He would of course never **dream of** playing these tapes”
- “All those mythical beasts your poets **dreamed up** in former centuries”
- “full of hideous **dreams from** which she struggled” *(COMPARA)*

METACHARACTERS

- **Empty brackets []**
  - allow 1 word to appear inbetween target elements

1 word between ‘lead’ (*v.*) and ‘to’

Query string: [lema="lead"][ ][word="to"]

- “half-glazed doors **led back to** the baking streets”
- “ the staircase that **led up to** the family flat ”
- “and here's a path that **leads straight to** it ”
- “we saw a track **leading off to** the left from the road” *(COMPARA)*

METACHARACTERS

- **Braces {}**
  - used to indicate the number of words permitted inbetween target elements
- **[ ]{n}**
  - indicates that exactly **n** words must occur between target elements

2 words between ‘have’ and ‘of’

Query string: [lema="have"][ ]{2}[word="of"]

- “but they **had a bit of** a bottleneck there at the time”
- “Analysis **has a way of** unravelling the self”
- “I like it -- it **has a touch of** class”

(COMPARA)

METACHARACTERS

- **[ ]{n,m}**
  - indicates that at least **n** but not more than **m** words must occur between target elements to match

1 or 2 or 3 words between ‘let’ and ‘down’

Query string: [lema="let"][ ]{1,3}[word="down"]

- “to **let the hair down** and put the knees up”
- “**letting the head hang down** like a bag”
- “You have to be strong, not **let things get you down**”

(COMPARA)

METACHARACTERS

- ❑ **Ampersand (&)**
  - ❑ used to combine attributes within one target element

“Fly” is a noun and a verb.

Query string: [lema="fly"&pos="V.*"]

searches for all forms of the word “fly” as a “verb”

- “**flew** out to join him on the first available plane”
- “matter of fact, I'm **flying** at their expense”
- “begin to **fly** through the air”