

Regular Expressions for Everyone



L I N G U A e m u n d i
LinguaeMundi



Inês Lucas

Quality Manager & Language Department Manager
LinguaeMundi

You **DO NOT**
need to

Know how to write regex

You **DO** need to

Identify situations where regex
may apply

Using regex, a case study: cleaning TMs

1

Unclean, polluted TMs

Control codes, entities and non-protected tags

Mojibake

Inconsistency in patterned information (dates, numbers with unit measures, hours)

Other problems that could be solved with memoQ standard features



2

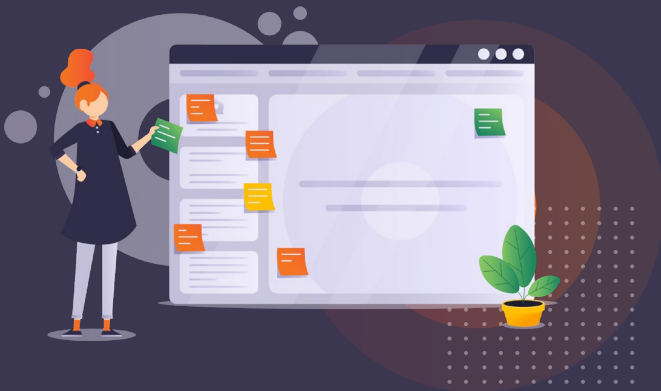
Unreliable matches

Inconsistent concordance results

Issues with pre-translated segments

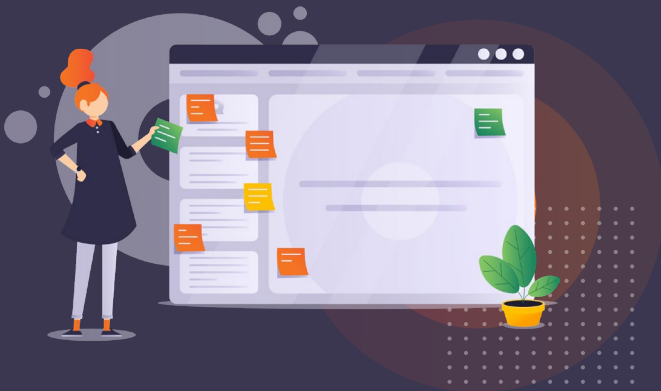
Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions
- 3 Train the language teams and create a checklist
- 4 Create the resources
- 5 Perform the cleaning
- 6 Import the clean translation memories



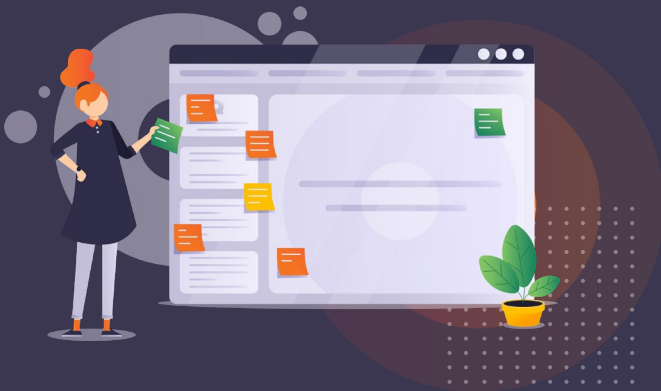
Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions**
- 3 Train the language teams and create a checklist
- 4 Create the resources
- 5 Perform the cleaning
- 6 Import the clean translation memories



Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions
- 3 Train the language teams and create a checklist**
- 4 Create the resources
- 5 Perform the cleaning
- 6 Import the clean translation memories



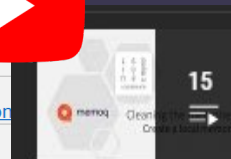
3

Train the language teams and create a checklist

| TM CLEANING CHECKLIST | | | |
|-----------------------|--|-------------|--|
| # | Activities / Tasks / Items | Status | Useful Link |
| 1 | Import the memory that was sent to you as a local resource | Completed | TM Cleaning 01 Create a local memory |
| 2 | Open the memory in the translation memory editor | Completed | |
| 3 | Remove duplicates with sources equal and same context | Completed | TM Cleaning 02 Translation Memory Editor |
| 4 | Remove duplicates with sources and targets both equal | Not Started | |
| 5 | Export the memory | Not Started | |
| 6 | Import the memory as a regular translation document to a local project in memoQ | Not Started | TM Cleaning 03 Import the memory as a translation |
| 7 | Import the regex library File named "TMClean Regex Library" | Not Started | TM Cleaning 05 Regex Assistant |
| 8 | Import the filter configurations as a local resource Files named "Control codes" and "Unprotected tags + entities" | Not Started | TM Cleaning 06 Regex Tagger |
| 9 | Leave any relevant comments throughout the cleaning, always tagging the relevant people if needed | Not Started | TM Cleaning 07 Comments |
| 10 | Lock rows in a different language using the lock/unlock tool | Not Started | |
| 11 | Filter for locked segments | Not Started | |
| 12 | Check the segments to see if memoQ's recognition is correct | Not Started | |
| 13 | Unlock any incorrectly flagged segment | Not Started | |
| 14 | Clean the filter | Not Started | |
| 15 | Use the regex "Find segments in other languages" to filter the source segments Select the relevant regex according to the source language | Not Started | |



YouTube playlist



Cleaning the memories

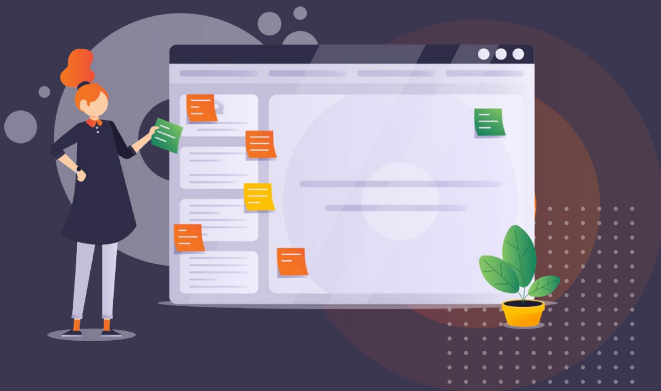
This playlist includes a process meant to clean the translation memories in memoQ. The videos were made to be watched in a certain order, that's why they're numbered. Have fun!

Checklist



Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions
- 3 Train the language teams and create a checklist
- 4 Create the resources**
- 5 Perform the cleaning
- 6 Import the clean translation memories



4

Create the resources

1

memoQ Regex Assistant library

2

Regex Tagger configurations for control codes, unprotected tags and entities

3

Customized QA settings

4

Test TM

5

TM Cleaning Checklist



4

Create the resources

- 1 memoQ Regex Assistant library
- 2 **Regex Tagger configurations for control codes, unprotected tags and entities**
- 3 Customized QA settings
- 4 Test TM
- 5 TM Cleaning Checklist



4

Create the resources

- 1 memoQ Regex Assistant library
- 2 Regex Tagger configurations for control codes, unprotected tags and entities
- 3 **Customized QA settings**
- 4 Test TM
- 5 TM Cleaning Checklist



4

Create the resources

- 1 memoQ Regex Assistant library
- 2 Regex Tagger configurations for control codes, unprotected tags and entities
- 3 Customized QA settings
- 4 **Test TM**
- 5 TM Cleaning Checklist



4

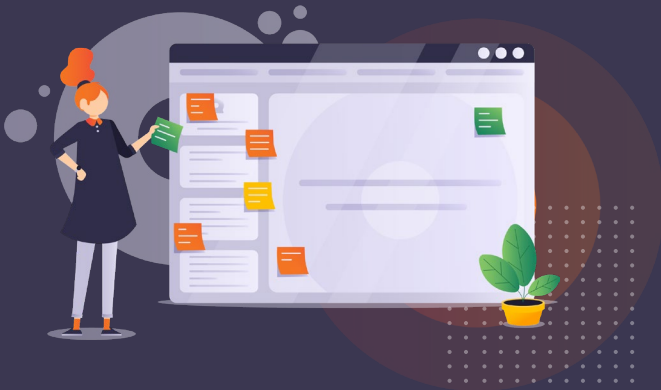
Create the resources

- 1 memoQ Regex Assistant library
- 2 Regex Tagger configurations for control codes, unprotected tags and entities
- 3 Customized QA settings
- 4 Test TM
- 5 **TM Cleaning Checklist**



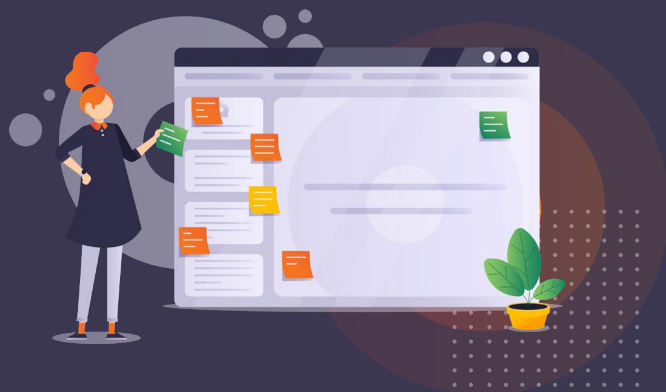
Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions
- 3 Train the language teams and create a checklist
- 4 Create the resources
- 5 Perform the cleaning
- 6 Import the clean translation memories



Our plan at LinguaeMundi

- 1 Identify patterns
- 2 Create and test regular expressions
- 3 Train the language teams and create a checklist
- 4 Create the resources
- 5 Perform the cleaning
- 6 **Import the clean translation memories**



Regex Assistant

Regex Assistant

Regex cheat sheet | Regex library

Find what:

Choose from the Regex cheat sheet or the Regex library above or start typing

Edit your Regex library | Add to Regex library

Insert regex to: Filter (target)

✓ Show replace options

Testing ground: No matches in testing ground

Copy some text here to test regex

Insert source segments | Insert target segments

Find

Regex Assistant

Regex cheat sheet | Regex library

Find what:

Choose from the Regex cheat sheet or the Regex library above or start typing

Replace with:

Edit your Regex library | Add to Regex library

Insert regex to: Advanced find and replace

⤴ Hide replace options

Testing ground: No matches in testing ground

Copy some text here to test regex

Insert source segments | Insert target segments

After replace:

Find and replace

Regex Assistant cheat sheet

| Regex Assistant | | Regex library | |
|-------------------|---|---------------|--|
| Regex cheat sheet | | | |
| . | Any character | \B | Not a word boundary |
| [0-9] | Any digit | ? | Occurs 0 or 1 times (quantifier) |
| \d | Any digit | * | Occurs 0 or more times (quantifier) |
| \D | Any character that is not a digit | + | Occurs 1 or more times (quantifier) |
| \w | Any digit or letter or underscore | {n} | Occurs exactly n times (quantifier) |
| \W | Any character that is not a digit, letter, or underscore | {n,m} | Occurs n to m times (quantifier) |
| \s | Any whitespace character (space, tab, line break) | | |
| \S | Any non-whitespace character | | |
| [a-z] | Any lowercase letter of the English alphabet | \P{L} | Any character that is not a letter |
| [^a-z] | Any character that is not a lowercase letter of the English alphabet | \p{L} | Any lowercase letter |
| [A-Z] | Any uppercase letter of the English alphabet | \P{Ll} | Any character that is not a lowercase letter |
| [^A-Z] | Any character that is not an uppercase letter of the English alphabet | \p{Lu} | Any uppercase letter |
| \p{L} | Any letter | \P{Lu} | Any character that is not an uppercase letter |
| | | [AbC] | Any character that is A, b, or C |
| | | [^AbC] | Any character that is not A, b, or C |
| | | (?i) | Case insensitive |
| | | (?m) | Multi-line mode. Treats rows as separate strings. |
| | | (?x) | Free-spacing mode. Allows using spaces, tabs and line breaks to make the regex easier to read. |
| | | ^ | Beginning of segment |
| | | \$ | End of segment |
| | | \b | Word boundary (beginning or end) |

Regex Assistant

Regex Assistant

Regex cheat sheet | **Regex library**

Find what:

Choose from the Regex cheat sheet or the Regex library above or start typing

Edit your Regex library | Add to Regex library

Insert regex to: Filter (target)

✓ Show replace options

Testing ground: No matches in testing ground

Copy some text here to test regex

Insert source segments | Insert target segments

Find

Regex Assistant

Regex cheat sheet | **Regex library**

Find what:

Choose from the Regex cheat sheet or the Regex library above or start typing

Replace with:

Edit your Regex library | Add to Regex library

Insert regex to: Advanced find and replace

⤴ Hide replace options

Testing ground: No matches in testing ground

Copy some text here to test regex

Insert source segments | Insert target segments

After replace:

Find and replace

Regex Assistant library

Search for a regex name or label

Built-in regexes:

- Date format conversion: (yy)yy. mm. dd. -> dd-mm-(yy)yy
- Date format conversion: (yy)yy. mm. dd. -> mm/dd/(yy)yy
- Date format conversion: dd-mm-(yy)yy -> (yy)yy. mm. dd.
- Date format conversion: dd-mm-(yy)yy -> mm/dd/(yy)yy
- Date format conversion: mm/dd/(y
- Date format conversion: mm/dd/(y
- Find acronyms (2+ all caps)
- Find alternative spellings of a word
- Find currency sign/code+space+va
- Find e-mail addresses
- Find numbers only segments
- Find proper nouns (except first wo

Search for a regex name or label

Your saved regexes:

- "Mas" at the beginning of a sentence
- Bibliography - Recognize single author publications
- Bibliography - Recognize single author publications
- Bibliography - Replace ", and" by ", e"
- Bibliography - Replace "In" by "Em"
- Exceptions to full written numbers 0-10, use numerals for the rest
- Find acronyms
- Find empty segments
- Find EN ordinal numbers - st, nd, rd, th
- Find entities starting with & and ending with ;
- Find hours with a colon and replace it with an I
- Find months with the first letter capitalized - P

TIP: use the library to add to your cheat sheet

Search for a regex name or label

Built-in

cheat

Your saved regexes:

- Generic negative look ahead
- Generic negative look behind
- Generic positive look ahead
- Generic positive look behind
- Non-breaking space

Labels:

- Cheat Sheet Look around
- Cheat Sheet Look around
- Cheat Sheet Look around
- Cheat Sheet Look around
- Cheat Sheet

Search for a regex name or label

biblio

Your saved regexes:

- Bibliography - Recognize single author publications followed by the year
- Bibliography - Recognize single author publications with two proper names followe...
- Bibliography - Replace ", and" by ", e"
- Bibliography - Replace "In" by "Em"

Labels:

- Bibliography References
- Bibliography References
- Bibliography References
- Bibliography References

Regex Assistant

Testing
ground

Regex Assistant

Regex cheat sheet

Regex library

Find what:
(?i)(\d[\d\p{Pd}\.,]*?)\s?(€|eur[os]*)

Replace with:
€\$1

Edit your Regex library

Insert regex to:

Hide replace options

Testing ground:
3€
12.50 EUR
9.99 euros
3.50–8.00 €
1,50 Euro

Insert source segments

After replace:
€3
€12.50
€9.99
€3.50–8.00
€1,50

Add to regex library

Give your regex a name. Add labels so that you can find the regex easier. Add a description so that you remember what the regex exactly does.

Name:
Change trailing euros to €XX

Labels (separated by commas):
number, currency

Description (shown in tooltip):
Changes euro currency expressions and dashed ranges to leading € format. Does not handle Swiss figures with apostrophes or number grouping with spaces. Case-insensitive.

Add to Library

Cancel

Description

Advanced Find and Replace

Regex Assistant

Regex cheat sheet

Regex library

?

Find what:

⌂

"(["]*)"

Replace with:

⌂

«\$1»

Edit your Regex library

Add to Regex library

Insert regex to: Advanced find and replace

⌆ Hide replace options

i

 Testing ground:

"The language barrier was overcome thanks to the skilled translator's assistance."

Insert source segments

Insert target segments

After replace:

«The language barrier was overcome thanks to the skilled translator's assistance.»

Auto-translation rules

| | | |
|-------------------------------------|---|-----------------------------------|
| OJ·A·111,·07.10.2010,· pp.·26-31 | 2 | JO·A°111·de· 07.10.2010,·p.º26 |
|-------------------------------------|---|-----------------------------------|

OJ·A·111,·07.10.2010,·pp.·26-31

JO·A°111·de·07.10.2010,·p.º26

| | | |
|----------|---|--------------|
| 134/2017 | 3 | n.ºº134/2017 |
|----------|---|--------------|

Decision·134/2017/EU

Decisão·n.ºº134/2017/UE

| | | |
|--------------|---|---------------|
| Article·1·bc | 3 | artigoº1.º-BC |
|--------------|---|---------------|

The employee handbook defines "misconduct" in Article 1·bc as any behavior that violates company policy and could result in disciplinary action.

O artigoº1.º-BC do manual do trabalhador define «má conduta» como qualquer comportamento que viole a política da empresa e que possa resultar em ações disciplinares.

Regex Tagger

Control
codes

Tag current document

Filter

Regex tagger

Filter configuration

Tags and entities [memoq.linguaemundi.pt]

Add cascading filter...

Remove

Rules

<[^\/*?> -> \$0

</[^\/*?> -> \$0

<[^\/*?> -> \$0

<[^\/*?> -> \$0

&[^\s]+?: -> \$0

Regular expression

<[^\/*?>

Tag type

☒ Open

☐ Close

☐ Required

Display text

\$0

Add

Change

Delete

☐ Rules handle tabs and newlines

Tags and
entities

Tag current document

Filter

Regex tagger

Filter configuration

Control codes - n.endash.rdblquote.rquote.r.ft [memoq.linguaemur]

Add cascading filter...

Remove

Rules

\\(n|endash|rdblquote|rquote|r|f|t) -> \$0

Regular expression

\\(n|endash|rdblquote|rquote|r|f|t)

Tag type

☐ Open

☐ Close

☒ Empty

☐ Required

Display text

\$0

Add

Change

Delete

☐ Rules handle tabs and newlines

Up •

Down •

Regex Tagger

Unprotected tags

Before

`<cf font="Verdana" size="10" complexscriptsize="10">`The equipment which is referred this manual is an instrumental rack.

`<cf font="Verdana" size="10" complexscriptsize="10">`O equipamento mencionado neste manual é um bastidor de instrumentação.

After

`<cf font="Verdana" size="10" complexscriptsize="10">` The equipment which is referred this manual is an instrumental rack.

`<cf font="Verdana" size="10" complexscriptsize="10">` O equipamento mencionado neste manual é um bastidor de instrumentação.

Control codes

Before

The conclusion from this is that the trade mark applied for should be rejected. `\rdblquote`

Termina concluindo que a marca registada deve ser recusada. `\rdblquote`

After

The conclusion from this is that the trade mark applied for should be rejected. `\rdblquote`

Termina concluindo que a marca registada deve ser recusada. `\rdblquote`

Quality Assurance

Edit QA settings

Segments and termsConsistencyNumbersPunctuationSpaces, capitals, charactersInline tagsLengthRegexSeverity

☒ Use regular expression checks

Warn if

forbidden regex match in target

Source regex

Target regex

[n|N]\.º(?!u00A0)

Correction

Description

Missing non-breaking space after n.º

☐ Expand tags to text before processing

Add

Update

Delete

"Regulation not properly translated" - missing regex replacement in target

"Forbidden pp." - forbidden regex match in target

"lowercase letter 'p' followed by a period, a space, a number, a hyphen, an

"Forbidden space before and after the hifen" - forbidden regex match in targ

"No space before and after the dash" - forbidden regex match in target

"Forbidden space before and after the forward slash" - forbidden regex mat

"Forbidden space after the minus, plus and plus-minus sign" - forbidden regex match in target

Missing non-breaking space after n.º" - forbidden regex match in target

"Capitalization issues - missing capital letter at the beginning of the item" - forbidden regex match in target

"Capitalization issues - forbidden capital letter at the beginning of the item" - forbidden regex match in target

"Segments can't start with isto or isso; alternative: Tal" - forbidden regex match in target

"In this case, numbers from 0 to 10 should be written in numeral form" - forbidden regex match in target

OK

Cancel

Help

Warnings

| Code | Description | Ignore |
|-------|--|--------------------------|
| 03206 | Missing non-breaking space after n.º - forbidden regex match in target N.º | <input type="checkbox"/> |

Close

Help

Filtering and sorting bar: source and target boxes

| Source | <div><div>Rx</div><div>Target</div></div> | <div><div>Rx</div><div></div></div> | <div><div></div><div></div><div></div><div></div></div> | Sort | No sorting |
|--------|---|--|---|------|------------|
| 316. | But normally for this the debt must be declared. | Mas, normalmente, para isso a dívida tem de ser declarada. | 0% | ✓ | |
| 6561. | But to do that, you need to have the right tools. | Mas para o fazer, precisa das ferramentas adequadas. | 0% | ✓ | |
| 6588. | But you need a customer relationship management system. | Mas precisa de um sistema de gestão de relações com os clientes. | 0% | ✓ | |
| 6681. | But how? | Mas como? | 0% | ✓ | |

Exporting the Regex library

```
C:\Users\ines-lucas\Desktop\TMCleanRegexLibrary.xml - Notepad++
Ficheiro Editar Procurar Visualização Codificação Linguagem Definições Ferramentas Macro Executar Plugins Janela ?
TMCleanRegexLibrary.xml x
25 .....<Item>CRLF
26 .....<Name>Find months with the first let
27 .....<FindRegex>Janeiro|Fevereiro|Março|A
28 .....<ReplaceRegex>/>CRLF
29 .....<Labels>CRLF
30 .....<Label>CRLF
31 .....<Value>TMClean</Value>CRLF
32 .....</Label>CRLF
33 .....<Label>CRLF
34 .....<Value>Old Spelling</Value>CRLF
35 .....</Label>CRLF
36 .....</Labels>CRLF
37 .....<Description>Following the new spell
38 .....</Item>CRLF
```

Exported library
.xml format

Convert the .xml
file to .html

FREEFORMATTER.COM

Search tools...

Free Online Tools For Developers

Buy me a coffee

Formatters

XML Formatter

JSON Formatter

HTML Formatter

SQL Formatter

Validators

XML Validator

JSON Validator

HTML Validator

XPath Tester

Credit Card Number Generator ...

Regular Expression Tester

Java Regular Expression Tester

Cron Expression Generator (Quartz)

Converters

XSD Generator

XSLT (XSL Transformer)

XML to JSON Converter

JSON to XML Converter

CSV to XML Converter

CSV to JSON Converter

YAML to JSON Converter

JSON to YAML Converter

Epoch Timestamp To Date

Encoders / Cryptography

XSL Transformer - XSLT

Converters / XSL Transformer - XSLT

This XSL Transformer (XSLT) let's you transform an XML file using an XSL (Extensible Stylesheet Language) file. You can also chose your indentation level if the result is an XML file.

The XSL Transformer fully supports XML namespaces, but the declarations MUST be explicit and MUST be on the root XML element of both your XML file and your XSL file. See the [XSLT Examples](#) section for details.

Option 1: Copy-paste your XML document here

Copy-paste your XML here

Option 2: Or upload your XML file

Escolher ficheiro

TMCleanRegexLibrary.xml

File encoding

UTF-8

Option 1: Copy-paste your XSL document here (Optional if XSD referred in XML using schemaLocation)

<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsl="http://www.w3.org/2001/XSL/Transform">
 <xsl:template match="/">
 <html>

Option 2: Or upload your XSL document

Escolher ficheiro

Nenhum ficheiro selecionado

File encoding

UTF-8

Exporting the Regex library

Find months with the first letter capitalized - PTpt

☐ TMClean ☐ Old Spelling

Find: Janeiro | Fevereiro | Março | Abril | Maio | Junho | Julho | Agosto | Setembro | Outubro | Novembro | Dezembro

Replace:

Description:

Following the new spelling agreement, months in Portuguese should not be capitalized

Regex Assistant library in .html format

Translating the Regex library

Project home TMCleanRegexLibrary.xml

Source Target

Sort No sorting

| | | | | |
|----|--|---|----|---|
| 6. | Find months with the first letter capitalized - PTpt | Encontrar meses com a inicial maiúscula – PTpt | 0% | ✓ |
| 7. | TMClean | TMClean | 0% | ✓ |
| 8. | Old Spelling | Antigo acordo | 0% | ✓ |
| 9. | Following the new spelling agreement, months in Portuguese should not be capitalized | De acordo com o novo acordo ortográfico, os meses em português não devem ser grafados com maiúscula | 0% | ✓ |

Changed ines-lucas 15/06/2023 00:24

View pane

```
</Item>
<Item>
  <Name>Encontrar meses com a inicial maiúscula – PTpt</Name>
  <FindRegex>Janeiro|Fevereiro|Março|Abril|Maio|Junho|Julho|Agosto|Setembro|Outubro|Novembro|Dezembro</FindRegex>
  <ReplaceRegex />
  <Labels>
    <Label>
      <Value>TMClean</Value>
    </Label>
    <Label>
      <Value>Antigo acordo</Value>
    </Label>
  </Labels>
  <Description>De acordo com o novo acordo ortográfico, os meses em português não devem ser grafados com maiúscula</Description>
</Item>
```

Translating the Regex library

Regex Assistant library in .html format translated into PT

Encontrar meses com a inicial maiúscula – PTpt

TMC

Finc

Rep

Des

De a

Edit regex

Name:

Encontrar meses com a inicial maiúscula – PTpt

Labels (separated by commas):

Antigo acordo, TMClean

Description (shown in tooltip):

De acordo com o novo acordo ortográfico, os meses em português não devem ser grafados com maiúscula

Find what:

Janeiro|Fevereiro|Março|Abril|Maio|Junho|Julho|Agosto|Setembro|Outubro|Novembro|Dezembro

Save

Cancel

agosto | Setembro | Outubro | Novembro | Dezembro

grafados com maiúscula

meses

Your saved regexes:

Encontrar meses com a inicial maiúscula – PTpt

Labels:

Antigo acordo TMClean

Regex Assistant library in .html format translated into PT





Regex Assistant library inside memoQ translated into PT

Translating the Regex library

Sort by label and export all the regex with the same label

Edit your Regex library

Filter for names or labels:
antigo

| <input checked="" type="checkbox"/> | Name | Labels | |
|-------------------------------------|---|------------------|---|
| <input checked="" type="checkbox"/> | Encontrar estações com a inicial maiúscula - PTpt | Antigo acordo +1 |   |
| <input checked="" type="checkbox"/> | Encontrar meses com a inicial maiúscula - PTpt | Antigo acordo +1 |   |

2 regexes selected

[Manage selected labels...](#) [Delete selected regexes](#) [Export selected regexes...](#) [Import library...](#)

Close

Why is this important?



Increased productivity



Fewer complaints



Lower stress levels



Time and cost efficiency



Reduce human error



Continuous improvement



Thank you!

Any questions?

L I N G U A
e m u n d i
LINGUAEMUNDI

ines@linguaemundi.pt