



Voice-text integrated system for interpreters

User Guide

Contenido

CORPUS	3
1. COMPARABLE CORPUS MANAGEMENT	3
1.1. <i>My corpora</i>	3
1.2. <i>Corpus query</i>	3
1.3. <i>Importing a corpus</i>	7
1.4. <i>Semi-automated corpus compilation</i>	8
2. PARALLEL CORPUS MANAGEMENT.....	8
2.1 <i>My corpora</i>	9
2.2 <i>Corpus query</i>	9
2.3 <i>Corpus importation</i>	10
GLOSSARY.....	11
3. DICTIONARY AND GLOSSARY MANAGEMENT	11
3.1. <i>My glossaries</i>	11
3.2. <i>Glossary management</i>	12
4. GLOSSARY QUERY	13
COMPLEMENTARY	14
5. TEXT SUMMARISATION SYSTEM.....	14
APPENDIX: MICROPHONE CONFIGURATION	15
APPENDIX: VERTICALIZED TEXT (.VRT)	17



CORPUS

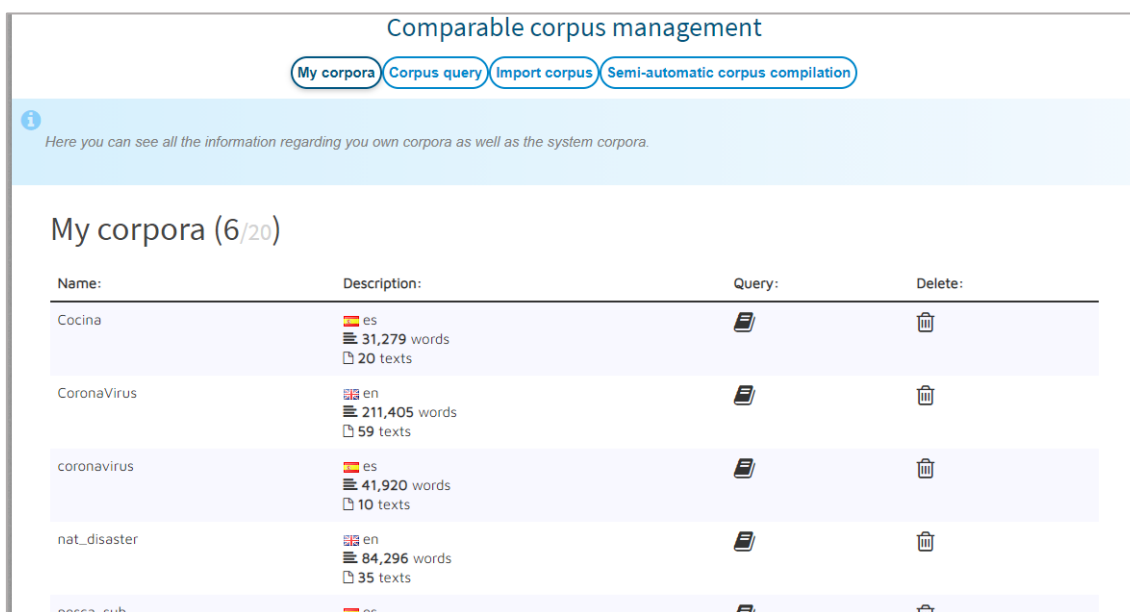
This module is designed to carry out the preparatory tasks for an interpretation, such as the creation and query of corpora, glossaries, etc. The different parts of this module are described below.

1. Comparable corpus management
















This module offers different functionalities related to the corpora. You can create your own corpus, either using text files or from an Internet search. Different types of queries can also be performed on the created corpus.

1.1. My corpora

This section shows the basic information for the corpora that have been created: name, language, number of words and number of texts that comprise it. It also allows you to delete a corpus by clicking on the  icon, or go to the corpus query by clicking on the  icon.



The screenshot displays the 'Comparable corpus management' interface. At the top, there are four navigation buttons: 'My corpora', 'Corpus query', 'Import corpus', and 'Semi-automatic corpus compilation'. Below this is an information icon and a message: 'Here you can see all the information regarding you own corpora as well as the system corpora.' The main section is titled 'My corpora (6/20)' and contains a table with the following data:

Name:	Description:	Query:	Delete:
Cocina	 es 31,279 words 20 texts		
CoronaVirus	 en 211,405 words 59 texts		
coronavirus	 es 41,920 words 10 texts		
nat_disaster	 en 84,296 words 35 texts		
nesca_sub	 es		

1.2. Corpus query

In this section, you can perform different queries over the corpora that you have created or imported. You just need to select one or several corpora from the list, type the search term and choose the type of query. To select several corpora, first select one and then press Ctrl+click on the next corpus that you would like to include in the selection.

In addition, several options can be set for the term, such as “Case sensitive” (makes the search case sensitive), “Diacritics insensitive” (ignore diacritics), “Grammatical category” (only terms matching the selected grammatical category will be shown). You can also choose if you want to perform the search by word form (as it appears in the corpus) or by lemma (lemmatized form of terms). The different types of queries that can be performed are described below.

- **Concordances**

Concordances are shown in a small context containing the given term in the selected corpora. The results can be sorted alphabetically according to the first word from the left (L1), the first word from the right (R1), the second word from the left (L2), the second word from the right (R2), and so on. In addition, several ordering levels can be set.

The screenshot displays the 'Corpus query' interface. It includes a list of corpora on the left with 'CoronaVirus_en' selected. The search term is 'disease'. Search options include 'Concordances' (selected), 'N-grams', 'Patterns', and 'Frequency list'. Ordering options include 'R1' (selected), 'L1-L5', and 'R1-R5'. A 'Delete' button is present. Below the query form, the results for 'CoronaVirus' are shown: '349 coincidences were found in 22 different texts (in 211,405 words [59 texts]; frequency: 1650.86 instances per million words)'. A 'Show/hide file names' button is also visible. The results list includes entries like '...es/ Diseases/ Coronavirus disease 2019 Coronavirus disease (COVID-19) Pandemic Public Advice Country & technical guidance Latest updates'.

- **N-grams**

In this case you must assign a value to N (2 for bigrams, 3 for trigrams, etc.). The N-grams containing the given term in the selected corpora are shown for the chosen N value.

Corpus query

Corpus name:

- cocina_es
- CoronaVirus_en
- medicina_es
- micorpus_es

Term:

Case sensitive

Diacritics insensitive

Grammatical category

Any ▼

Word form

Lemma

Search type:

Concordances [?](#)

N-grams

Patterns

Frequency list [?](#)

Add selected terms to:

Results for corpus: **CoronaVirus**

Mark all

N°	Resultados	N° appearances	Percent
<input type="checkbox"/> 1	Coronavirus disease 2019	91	13.48% (total: 0.043%)
<input type="checkbox"/> 2	for Disease Control	26	3.85% (total: 0.0123%)
<input type="checkbox"/> 3	Disease Control and	24	3.56% (total: 0.0114%)

- **Patterns**

The terms or collocations that match the chosen pattern and contain the given term in the selected corpora are displayed. Custom patterns can also be created. If the “Ignore term” option is checked, the term will not be considered.

Corpus query

Corpus name:

Term:

Search type:
 Concordances
 N-grams
 Patterns
 Frequency list

Patterns:
 Custom

 Ignore term

Case sensitive
 Diacritics insensitive

Grammatical category...
 Word form
 Lemma

Add selected terms to:

Results for corpus: **CoronaVirus**

Select all

N°	Results	N° appearances	Percentage
<input type="checkbox"/> 1	infectious disease	16	47.06% (total: 0.0076%)
<input type="checkbox"/> 2	Respiratory disease	5	14.71% (total: 0.0024%)
<input type="checkbox"/> 3	Cardiovascular disease	3	8.82% (total: 0.0014%)

- Frequency list

This functionality offers a list of the most frequent words that could be candidates for terms or phraseological units in the selected corpora. You do not need to enter any term, only the grammatical category you do not want to be shown ("Noun", "Adjective", "Verb", "Preposition", "Conjunction", "Adverb".)

Corpus query

Corpus name:

Term:

Search type:
 Concordances
 N-grams
 Patterns
 Frequency list

Stop categories:
 Preposition
 Verb
 Adjective
 Noun

Case sensitive
 Diacritics insensitive

Grammatical category

Word form
 Lemma

Add selected terms to:

Results for corpus: **CoronaVirus**

Mark all

N°	Resultados	N° appearances	Percent
<input type="checkbox"/> 1	New	3231	6.81% (total: 1.5283%)
<input type="checkbox"/> 2	Cases	2203	4.64% (total: 1.0421%)
<input type="checkbox"/> 3	source	2161	4.55% (total: 1.0222%)
<input type="checkbox"/> 4	Deaths	1054	2.22% (total: 0.4986%)
<input type="checkbox"/> 5	COVID-19	770	1.62% (total: 0.3642%)

After performing the query, you can select (by checking the checkbox) the resulting terms (except for *Concordances*) and add them to a glossary that you can later consult and edit (see section *Glossary query*). To do this, select a glossary from the drop-down menu or choose a name to create a new one and press the "Add" button. In order to quickly select several terms, you have two options:

- 1) Check "Select all" to select all terms.
- 2) Select a range. To do this, check an initial checkbox and, while pressing the SHIFT button, check an ending checkbox. All checkboxes in between will be automatically checked.

The screenshot shows a search results interface. At the top, there are 'Search' and 'Clear' buttons. Below them is a section 'Add selected terms to:' with a dropdown menu for 'Select glossary...' and an 'Add' button. The dropdown menu is open, showing options like 'Pescas submanna', 'Natural disasters', 'nat_disasters', 'Covid-19', 'teletrabajo', 'coronavirus', 'Glossary WORKSHOP', 'ener_glo', 'miglosario', and '+ Create Glossary'. Below the dropdown, there is a table of results for a corpus. The table has columns for 'N°', 'Results', 'N° appearances', and 'Percentage'. The results are as follows:

N°	Results	N° appearances	Percentage
<input type="checkbox"/> 1	infectious diseases	16	47.06% (total: 0.0076%)
<input type="checkbox"/> 2	Respiratory disease	5	14.71% (total: 0.0024%)
<input type="checkbox"/> 3	Cardiovascular disease	3	8.82% (total: 0.0014%)

1.3. Importing a corpus

In this section you can create your own corpus from a set of texts. To do so, choose a name for the corpus and select the text files from your computer. These files can be in TXT or VRT¹ (VeRTicalized Text) format. Select the language of the texts. After pressing the "Import" button, the corpus will be created, and you will be able to consult it in the section *Corpus query*.

The screenshot shows the 'Comparable corpus management' interface. At the top, there are four tabs: 'My corpora', 'Corpus query', 'Import corpus', and 'Semi-automatic corpus compilation'. Below the tabs, there is a message: 'Create a corpus from your own text files. Type a name, upload your files and select the language.' The main form is titled 'Import corpus' and contains the following fields:

- Corpus name:** A text input field.
- Files:** A button labeled 'Add'.
- Select language:** Radio buttons for 'Spanish' (selected) and 'English'.
- Import:** A green button.

¹ The columns order must be: token, part of speech, lemma. See *Appendix: Verticalized text*.

1.4. Semi-automated corpus compilation

In this section you will be able to create a corpus using texts from the Internet. You just need to type a search term or phrase (you can use the usual search techniques such as putting the phrase in quotation marks to search for an exact match or typing the "-" Boolean operator before a word to exclude it from the search and select a language. After pressing the "Search" button, a list of related websites will be displayed, from which you will be able to choose as many as you consider appropriate to add to your corpus. Then you only have to choose a name for the corpus and select one of the three following options: "Import and download .txt" (the corpus will be downloaded in plain text format), "Import and download .vrt" (the corpus will be downloaded in vertical text format with morphological labelling, see *Appendix: Verticalized text*) or "Import only". In any case, a corpus which can be consulted in the section *Corpus query* will be created. You can also choose whether to combine all the obtained texts into a single file (by selecting "Create a file with all the information") or create a text file for each chosen web resource in addition to a metadata file (by selecting "Create a file for each web").

The screenshot shows a web interface for creating a corpus. At the top, there is a search bar with the text "lockdown", a "Select search engine" section with radio buttons for Level 1 (selected), Level 2, Level 3, Level 4, and Level 5, and a "Select language" section with radio buttons for Spanish and English (selected). A green "Search" button is located below these options.



Below the search bar, there are "100 results" and a control bar with "Select all", "Remove all", a text input "n", and a "Select" button. The results are displayed as a list of cards, each with a title, a snippet of text, and a "Go to website" button. The first card is titled "National lockdown: Stay at Home - GOV.UK" and contains the text "04/01/2021 · Summary: what you can and cannot do during the national lockdown. You must stay at home. The single most important action we can all take is ...". The second card is titled "Prime Minister announces national lockdown - GOV.UK" and contains the text "04/01/2021 · The Prime Minister has announced a national lockdown and instructed people to stay at home to control the virus, protect the NHS and save lives. The decision follows a rapid rise in infections ...". The third card is titled "lockdown - latest news, breaking stories and comment - The ..." and contains the text "lockdown. Voices. Victoria Richards A child used an app to tell us his mother died – it was devastating. Home News. Welsh schools could reopen from 22 February in plan to ease lockdown. Features ...". There is an "Add URL" button at the bottom right of the results area.

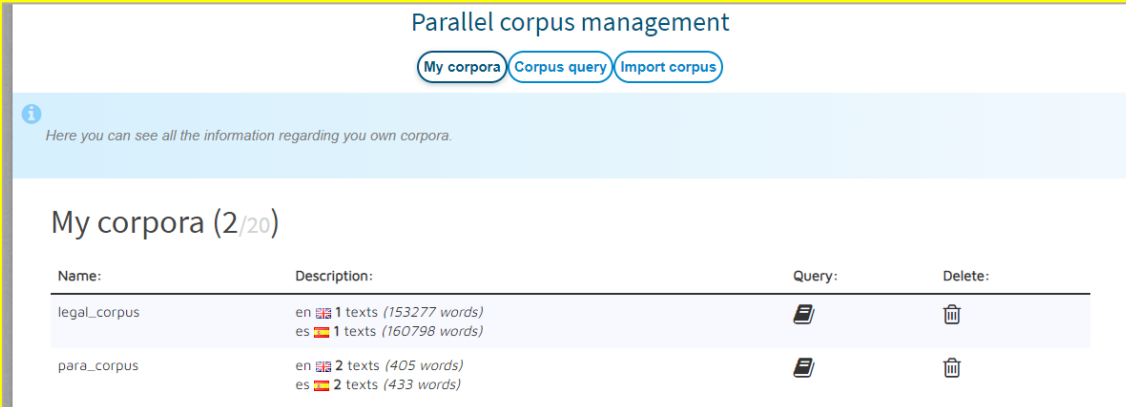
On the right side, there is a dashed box containing an "Added URLs (2)" section with two URLs: "https://www.gov.uk/guidance/nation..." and "https://www.gov.uk/government/ne...". Below this is a "Corpus name" input field with the text "Name". Underneath is a section titled "What do you want to do?" with three radio button options: "Import and download .txt", "Import and download .vrt", and "Import only" (selected). Below this is a "How?" section with two radio button options: "Create one file with all information" and "Create one file per web" (selected). A green "Process" button is located at the bottom right of the dashed box.









2. Parallel corpus management

In this module you can create your own parallel corpus using text files. Different types of queries can also be performed on the created corpus.

2.1 My corpora

This section shows the basic information for the parallel corpora that have been created: name, language, number of words and number of texts that comprise it. It also allows you to delete a corpus by clicking on the  icon, or go to the corpus query by clicking on the  icon.



Name:	Description:	Query:	Delete:
legal_corpus	en  1 texts (153277 words) es  1 texts (160798 words)		
para_corpus	en  2 texts (405 words) es  2 texts (433 words)		

2.2 Corpus query

In this section, you can perform different queries over the parallel corpora that you have created or imported. You just need to select one or several corpora from the list and type the search term. To select several corpora, first select one and then press Ctrl+click on the next corpus that you would like to include in the selection.

In addition, several options can be set for the term, such as “Case sensitive” (makes the search case sensitive), “Diacritics insensitive” (ignore diacritics), “Grammatical category” (only terms matching the selected grammatical category will be shown). You can also choose if you want to perform the search by word form (as it appears in the corpus) or by lemma (lemmatized form of terms).

After clicking “Search” button, fragments matching the typed text in selected corpus will be shown next to the equivalent in the other language corpus fragments.

Corpus query

Corpus name:

legal_corpus [en][es]

para_corpus [en][es]

Term:

budget

Language:

ES
 EN

Case sensitive
 Diacritics insensitive

Grammatical category... ▾

Word form
 Lemma

Search

Clear

legal_corpus: 69 results

EN	ES
Article 17 The budget of the Union for the financial year 2004 shall be adapted to take into account the accession of the new Member States through an amending budget taking effect on 1 May 2004.2 .	Artículo 17 El Presupuesto de la Unión para el ejercicio presupuestario 2004 será adaptado , a fin de tener en cuenta la adhesión de los nuevos Estados miembros , mediante un presupuesto rectificativo que entrará en vigor el 1 de mayo de 2004.2 .
Article 17 The budget of the Union for the financial year 2004 shall be adapted to take into account the accession of the new Member States through an amending budget taking effect on 1 May 2004.2 .	Artículo 17 El Presupuesto de la Unión para el ejercicio presupuestario 2004 será adaptado , a fin de tener en cuenta la adhesión de los nuevos Estados miembros , mediante un presupuesto rectificativo que entrará en

2.3 Corpus importation

In this section, you can create your own parallel corpus from a set of text files. The name of these files must be in a specific format: the name of the files which are equivalent in each language must be the same except for the final part of the name, which must include the language code preceded by a low dash (_xx, being "xx" the language code. For example, the Spanish file filename1_es.txt corresponds to the English file filename1_en.txt, the Spanish file file_name2_es.txt corresponds to the English file file_name2_en.txt, etc.). Then, choose a name for the corpus and select the text files from your computer. These files can be in TXT or VRT (VeRticalized Text) format. After pressing the "Import" button, the corpus will be created, and you will be able to consult it in the section Corpus query.

Parallel corpus management

[My corpora](#)
[Corpus query](#)
[Import corpus](#)

Create a corpus from your own text files. Type a name and upload your files.

Import corpus

Corpus name:

Files:

The name of the files which are equivalent in each language must be the same except for the final part of the name, which must include the language code preceded by a low dash (_xx, being "xx" the language code). Example:

Files of the Spanish corpus: *filename1_es.txt; file_name2_es.txt*
Files of the English corpus: *filename1_en.txt; file_name2_en.txt*

GLOSSARY

3. Dictionary and glossary management

In this section, you can see and manage the glossaries you have created.

3.1. My glossaries

Here you can see basic information about the glossaries you have created (name, description, number of terms) as well as create and delete glossaries. In order to create a new glossary, you just need press the "Create" button **and type the glossary information (name, languages and description)**. In doing so, an empty glossary will be created. You can change the description of any glossary by pressing on the edit icon which appears in the description when you hover your mouse over it. Then you can either confirm the change or discard it . To delete a glossary, press on the corresponding delete icon: . In order to add terms to a certain glossary you must go to *Glossary management* section (instructions are in the next section). If you click on the icon, you will be automatically redirected to the management of the selected glossary.

Glossary management

My glossaries
Glossary management






i Here you will be able to see the information regarding your glossaries. You can also create a new one or delete an existing one.

My glossaries (9/20)

+ Create

Name:	Description:	Manage:	Delete:
Pesca submarina	☰ 21 terms Glosario de pesca	☰	🗑️
Natural disasters	☰ 24 terms Glossary about natural disasters terms	☰	🗑️
nat_disasters	☰ 6 terms Natural disasters glossary	☰	🗑️
Covid-19	☰ 104 terms	☰	🗑️

3.2. Glossary management

In this section you can edit the created glossaries (modify, add, etc.). Just select a glossary from the list and click on "Show". To edit any term in the glossary, press on it or on the edit icon  which appears when you hover the mouse cursor over it. Then you can either confirm the change (pressing  or ENTER) or discard it (pressing  or ESC). If you click on the external resources search icon , a drop-down list with different resources (Google, Wikipedia, Linguee, etc.) will be displayed. When selecting any of these resources, a search for the selected term in that resource will be performed. You can also delete a glossary entry by pressing on the corresponding delete icon , or add a new blank one by pressing the "Add" button.

Moreover, the tool allows both importing terms into glossaries and exporting those glossaries. To import terms to a glossary, press the "Import" button and select the desired file (the allowed formats are XLSX, XLS, CSV and ODS). **The first row of this file must include the language codes (en, es, fr...).** This way all the terms included in your file will be copied to the current glossary. To export a glossary to an XLS file, just click the "Export" button and the file will be automatically downloaded.

Glossary management

Glossary name:
 Glosario de prueba ▾

Show Clear

ES EN

Prueba trial

+ Add

Export | Import ▾

File formats: [XLSX](#) [XLS](#) [CSV](#) [ODS](#)

Seleccionar archivo Ningún archivo seleccionado Import terms

4. Glossary query

In this section you can use the glossaries previously created. You just have to select the desired glossaries, press on "Load Glossaries" and a search field will appear. When typing a term, the results will start appearing automatically in a quick and convenient way. If you check the "Fuzzy match" option, results with partial coincidences will be shown too. In addition, you can perform the search using your voice, by clicking the icon or pressing the "s" key.

Pesca submarina
21 terms
Glosario de pesca

Natural disasters
24 terms
Glossary about natural disasters terms

nat_disasters
6 terms
Natural disasters glossary

Covid-19
104 terms
Coronavirus glossary

teletrabajo
11 terms
Glosario sobre el trabajo a distancia

coronavirus
10 terms
No description

Glossary WORKSHOP
78 terms
No description

ener_glo
10 terms
No description

miglosario
8 terms
No description

Load glossaries
✔ Covid-19 (104)
✔ coronavirus (10)

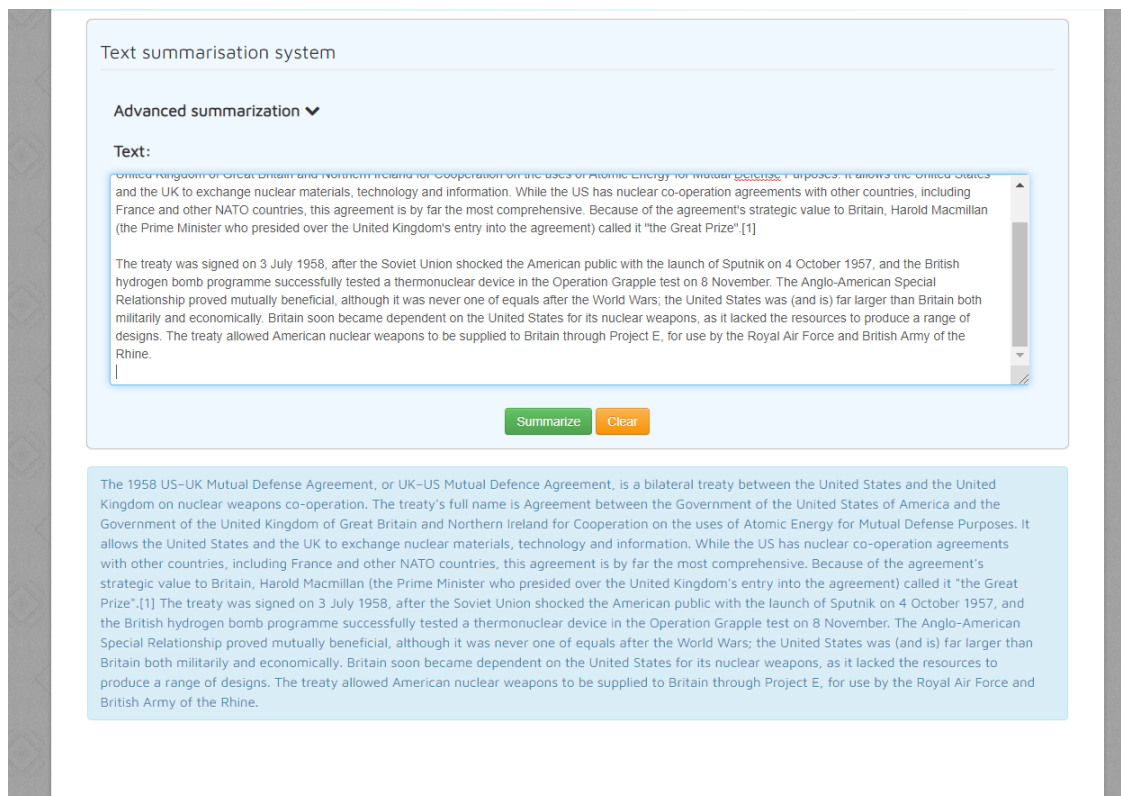
Search: Fuzzy match

6 results.

ES	EN
días	days
enfermedad	disease
disponible	available
diario	daily
diferente	different
murio	died

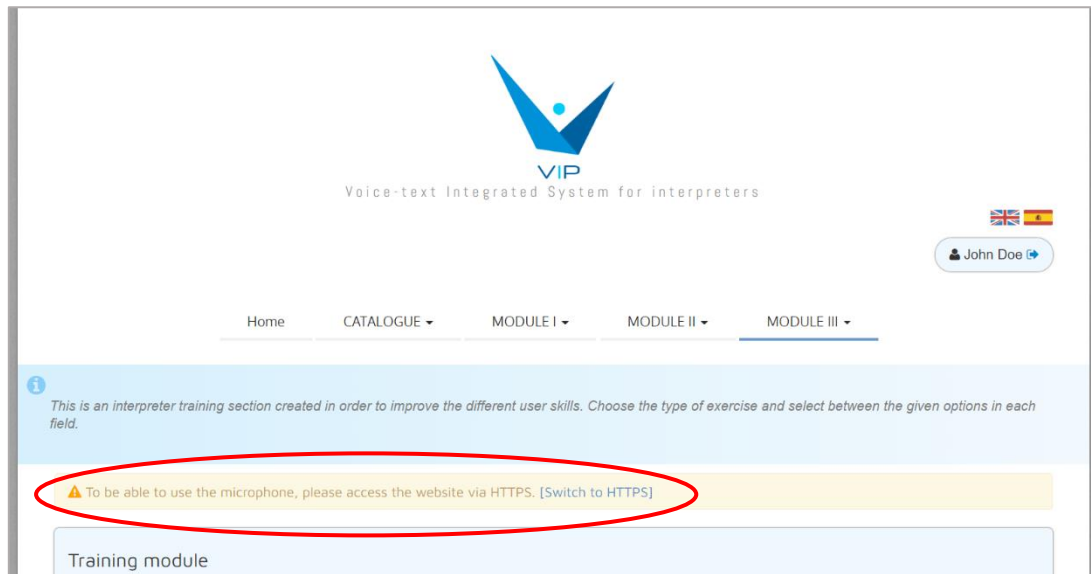
5. Text Summarisation system

In this section you can summarize a text. For this purpose, you can either load one or several plain text files (.txt), paste the text directly into the corresponding field, or use the text extraction tools to get the text from an URL or a PDF file. These tools will show the text in the corresponding field and allow the user to download this text in .txt format. Then, choose the type of summary by selecting the number of words or the percentage of the original text that it will take up. After pressing the "Summarize" button, the result will be shown next to the original text.

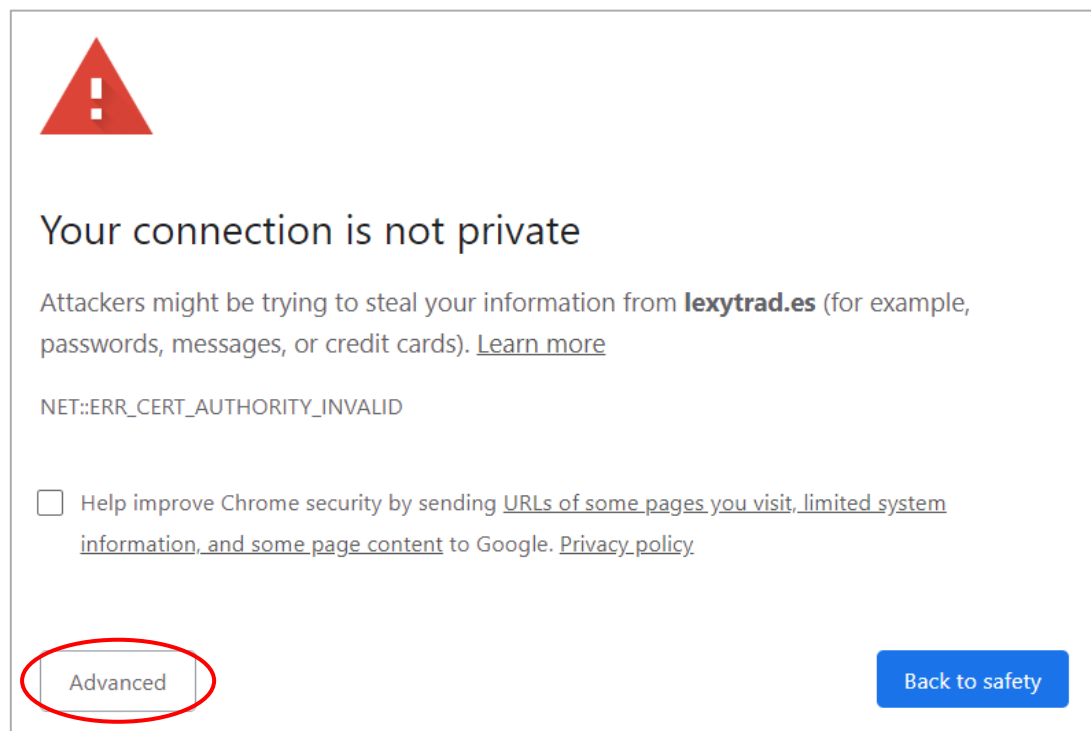


Appendix: Microphone configuration

In order to use the microphone in the system, you must access the platform via the secure HTTPS protocol. Otherwise, if you are accessing via HTTP, the following warning message will appear in the sections where the use of the microphone is required:



To access via HTTPS, click on "Switch to HTTPS" and the browser will display the following warning message:



To continue, press the "Advanced" button and the following message will be displayed:

This server could not prove that it is **lexytrad.es**; its security certificate is not trusted by your computer's operating system. This may be caused by a misconfiguration or an attacker intercepting your connection.

[Proceed to lexytrad.es \(unsafe\)](#)

Finally, click on "Access lexytrad.es (non-secure site)" and you will return to the system, where you will be able to use the microphone.

Appendix: Verticalized text (.vrt)

Verticalized text (.vrt), also known as one-word-per-line, is a format in which each token is in a different line. It can contain additional information related to each token, as the grammatical category or lemma (separated by tabulations). Structures such as sentences or paragraphs can be set too using XML tags.

If we take this text as example:

This is the first sentence. This is the second.

Verticalized text would be:

```
This
is
the
first
sentence
.
This
is
the
second
.
```

We can include additional information (such as grammatical category or lemma) next to each token, leaving one tabulation separation:

```
This      DET      this
is        VER      be
the       DET      the
first     ADJ      first
sentence  NOU      sentence
.         PUN      .
This      DET      this
is        VER      be
the       DET      the
second    ADJ      second
.         PUN      .
```

In addition, we can set structures such as sentences (<s>) or paragraphs (<p>) by using XML tags:

<p>		
<s>		
This	DET	this
is	VER	be
the	DET	the
first	ADJ	first
sentence	NOU	sentence
.	PUN	.
</s>		
<s>		
This	DET	this
is	VER	be
the	DET	the
second	ADJ	second
.	PUN	.
</s>		
</p>		